(ТИТУЛ)

Асқар Құдайбергенұлы ЖҰБАНОВ

ҚАЗАҚ ТІЛ БІЛІМІ: ҚОЛДАНБАЛЫ ЛИНГВИСТИКА

*** * ***

КАЗАХСКОЕ ЯЗЫКОЗНАНИЕ: ПРИКЛАДНАЯ ЛИНГВЫИСТИКА

Баспаға ҚР БжҒМ ҒК А.Байтұрсынұлы атындағы Тіл білімі институтының
Ғылыми кеңесі ұсынған

Шығарушылар алқасы:

А Н Н О Т А Ц И Я

(и т. д.)

ҚҰРМЕТТІ ОҚЫРМАН!

Оқырман қауымына ұсынып отырған «Қазақ тіл білімі: қолданбалы лингвистика» атты моногафияда 1967 жылдан 2012 жыл аралығындағы ғылыми басылымдарда автордың жарияланған мақалалары мен халықаралық, республикалық деңгейде өткен ғылыми-теориялық және тәжірибелік конференциялар мен семинарларда жасаған баяндамаларының тезистері көрініс тапты. Мақалалар реті хронологиялық тәртіппен берілгендіктен қатаң тақырыптық тараушаларға бөліп, шектеу әрекеті қиындық туғызды. Сондықтан оларды тақырыптық топқа бөлу принципін жалпы түрде: 1) автордың аспирантурада оқып жүрген ғылым жолындағы алғашқы кезеңі – кандидаттық диссертация жазу мен қорғау кезеңі (1973 ж.): I Бөлім – «Қазақ мәтінінің квантитативті құрылымына» қатысты мақалалар; екінші кезең – докторлық диссертация жазу мен қорғау кезеңі (2002 ж.): II Бөлім – «Қазақ мәтінінің мазмұнынын формалдау»; 3) үшінші кезең – соңғы кездердегі компьютерлік лингвистика саласына қатысты: III Бөлім – «Қазақ тілінің «Тіл – қазына» атты ұлттық компьютерлік қорын» құруға қатысты және 4) төртінші кезең – IV Бөлім: «Қазақ тілінің ұлттық мәтіндер корпусын құруға» қатысты мәселелер жайында жазылған мақалалар қамтылды. Бұл жерде ескерте кететін жайт – кейбір мақалалардың жоғарыда аталған бөлім тақырыптарына тікелей қатысты болмауы.

Қорыта айтқанда, автордың мақалалар жинағы ретіндегі «Қазақ тіл білімі: қолданбалы лингвистика» атты монографияда қамтылған мақалалар жиыны автордың 75 жасқа толу мерейтойы кезіндегі шәкірттерінің ұстазына жасаған сый-құрметі деп білген жөн сияқты.

АЛҒЫ СӨЗ

(Асқар Құдайбергенұлы Жұбанов – 75 жаста)

Күздің бір қара суық күнінде – 13 қарашада жарық дүниеге іңгәлап бір нәресте келді. 19 қараша күні перзентханадан алып келген баласын әкесі құшырлап иіскей берген сәтте сол кезеңнің әзірейілдеріндей болып көрінетін НКВД адамдары сау ете қалды. Үй ішін тас-талқан етіп тінтіп, қағаз біткеннің барлығын жинап алып, алмағанын жыртып, алды 11-12-лерде соңы алты күндік алты баланың үрейін ұшырып, аналарын көз жасына буып, үй иесін ала жөнеледі.

Оқиға мекені – Алматы қаласы, уақыты – 1937 жыл еді, ал шаңырақ иесі – сол кездің өзінде Одаққа аты танымал түркітанушы әріптестерімен терезесі тең тұрған ғалым, профессор Құдайберген Жұбанов болатын.

Биыл, 2012 жылдың қараша айында еліміздің тіл мамандары, магистранттар мен PhD докторонттар, студенттер мен университет оқытушылары, бүкіл зиялы қауым болып, атын үлкен құрметпен сүйіне атап, еңбегін елге паш етіп бір ғалымның мерейтойын өткізгелі отыр. Той иесі – баяғы әкесінен алты күндігінде қалып, заман мен өмірдің небір қырларын өтіп, еңбегімен, талабымен, таланытымен көрінген жақсы ғалым үлгілі ұстаз, филология ғылымдарының докторы, А.Байтұрсынұлы атындағы Тіл білімі институтының бас ғылыми қызметкері, профессор Асқар Құдайбергенұлы Жұбанов!

Тек бір отбасы өмірінде ғана емес, бүкіл қазақ халқының қиын да қилы тірлік-болмысында есте қалатын бұл екі оқиғаның аралығында өткен 75 жыл қасқағым сәттей. Бұл – бір қарағанда. Екінші қырынан үңіле көз жүгіртсек, қаншама ауыр жүк арқалаған жылдар, жан таласа өткен іс-қарекеттер күнделікті тұрмыста бастан кешкен тапшылық пен талпыныс, қаза мен қайғылы тұстар өтпеді дейсің?!

Асқар бір жасқа толмай-ақ «жер аударылып», туған қаласы Алматыдан әке-шеше отаны – Ақтөбе өлкесінен бір-ақ шығады. Дәлірек айтсақ, 1922 жылы «Қызыл Отау» деген жаңа орында Құдайберген жігітпен некесі қиылған үлбіреген ақ құбаша сұлу жас қыз Раушан енді алты баланың анасы болып толыға түскен шағында «халық жауының» әйелі болып шыға келеді. Әрі қарай өзі де ұсталып, лагерьлерге айдалудан, балаларының «детдом» деген аты суық үкімет құзырындағы үйлерге таратылып кетуінен сақтанып, алты баланы Алматыдан ала жөнелген Раушан жеңгейдің басына ауыртпалық мықтап түседі. Колхоздың қара жұмысына жегіліп, ауыр қайғы-қасірет шегіп жүргенде сүті қайтып кетеді. Асқар ана сүтін де еме алмай, кәдімгі резеңкі еміздікті де көрмей (сол кезде ауылда емізік қайдан болсын), сиыр сүтін қасықпен жұтып, нәресте шағын өткізді. Сәби Асқарды апасы (әпкесі) 12-13 жастағы Мүслима, 8-10 жастардағы ағалары Ақырап пен Есет «арқалап» өсіреді, ойнатады, жұбатады, еркелетеді. Өздері де балалық шақтан озбаған «тәрбиешілерінің» оны ойыншық көріп (Мүслиманың сөзі) жұлқылағандары ананың аялы алақанын, әкенің иығына мінгізіп, төбесіне шығарғанын жоқтатпағандай еді.

Сөйтіп, Асқар өсе бастайды. Мектепке барады. Бірақ мектеп жылдары да біркелкі, бір жерде өтпейді. Ауылда да, қалаларда да оқиды. Әсіресе апасы Мүсилима институтты бітіріп, Петропавл қаласына жұмысқа барғанда, апасы мен іні, сіңлілерін көшіріп алады. Асқар Петропавлдағы қазақ мектебіне барады, бірақ ғажабы – бұл мектеп «қазақ мектебі» аталғанымен қоғамдық пәндер қазақ тілінде оқытылып, ал математика, физика, химия сабақтарын орыс мұғалімдері орыс тілінде оқытады. «Медальдың екі жағы болады» деп орыс ағайындар айтқандай, мұндай «аралас тілді» мектептің пайдасы да тиді – дейді Асқар. Осы сабақтарды орысша оқудың нәтижесінде ол орыс тілін де жақсы меңгеріп алады. Ал физика, математика пәндері үшін орысша жақсы білудің күні бүгінге дейін қажет болып отырғандығын жасырмайық, өйткені оқу орындарында оқулықтар мен қосымша құралдардың көбі орыс тілінде екендігі, оқытушылардың да көбі (қазағы да) орыс тілділер екендігі мәлім. Осы себептен бе, әлде адамның өзіне белгісіз бір нәрсеге бүйрегі бұрып тұратын «құдайдың құдіретінен» бе, әйтеуір Асқар физика пәнін жақсы көреді, радиотехникамен шұғылданып, қысқасы, техникалық білім алуды қалайды. Тау-кен институтына түспек болып емтиханды ол жақсы тапсырады, бірақ… пайдалы қазбалар, табиғи байлықтар ол кезде стратегиялық мәні бар болғандықтан, бұған «халық жауларының ұрпақтары» қабылданбайды. Ағасы Есеттің кеңесімен бір кезде әкелері жұмыс істеген Абай атындағы ҚазПИ-дің математика факультетіне түседі. Жақсы бітіреді. Алматыдағы Мемлекеттік ауылшаруашылық институтына «жоғарғы математика» кафедрасының ассистенті, оқытушысы болып жұмысқа кіріседі. Онда 7 жыл істейді.

Кейде бір мақсатты ойлауын адам ойлайды, ал тағдыр (жазмыш) өз дегенін істеттіреді ғой, немесе кей-кейде, «кездейсоқтық» атты мықты құбылыс күш көрсете ме, Асқар кенет қазақ тілін математикалық әдіспен, оның ішінде математикалық статистика мен ықтималдылықтар теориясы әдістерімен зерттеу жолына бет бұрады. Бұған Алматыда өткен бір ғылыми конференцияның және оған келген «Сөз статистикасы» (Статистика речи) атты Бүкілодақтық топты басқаратын атақты профессор Раймонд Генрихович Пиотровский мен математик қазақ ғалымы Қалдыбай Бектаевтармен кездесіп, әңгімелескені себепкер болады. Кездейсоқ конференция, кездейсоқ жүздесулер Асқардың ғылымдағы жолын бұрып әкетеді. Бүгінгі қазақ тілтаным саласында Асқар Жұбанов еңбектерінің орны айрықша деп ғалымның 75 жылдығына орай оны марапаттап, жоғары бағалап отырғанымыздың себебі – Асқар Жұбанов тіл білімінің бүгінгі таңдағы маңызы зор, аса қажет жаңа саласын бастаушылардың бірі ғана емес, оны әрі қарай алып барушы нағыз маман екендігі. Бұл салалар – қолданбалы лингвистика, компютерлік лингвистика. Қазақ тілін қолданбалы тіл білімінің статистика, формалды модельдер, мәтін лингвистикасы атты салаларына қатысты зерттеу, бұлар бойынша оқу құралдарын жазу – ғалымның зор табысы. Бұл табыс осы оқулықтардың қазақ тілінде жазылғандығымен де бағасы арта түседі. Көпшілігіміз жете түсіне бермейтін формалды логика дегенді тіл ғылымына қарай бұру, тілді (атап айтқанда, қазақ тілін) зерттеуде

математикалық әдістерді қолдану, соның нәтижесінде «Математикалық лингвистика», «Статистикалық лингвистика», «Компьютерлік лингвистика», «Формалды модельдер» деп аталатын ғылым салаларын қазақ тілін зерттеуге қажетті пәндер деп, өмірге әкелгендігімен де Асқар Жұбановтың еңбегі жоғары бағаланады. Бұл оқу құралдары (қазақ тіліндегі тұңғыш оқулықтар десек те болады) филология факультеті студенттері мен магистранттарына қолданбалы лингвистика мамандарына, компьютер мүмкіндіктерін қоғамдық ғылымдар салаларына, оның ішінде қазақ тіл білімінде қолданумен айналысатын мамандарға арналған. Ең бір құндылығы – бұл оқу құралдарының бүгінгі күнде оқу процесінде қолданылып жатқандығы. Осы пән салаларынан профессор Асқар Жұбановтың бірнеше жылдан бері әл-Фараби атындағы Қазақ ұлттық университеті мен Абай атындағы Қазақ ұлттық педагогикалық университеті Магистратура және PhD докторантура институтының магистранттары мен PhD докторанттарына дәріс оқып, жастарды тәрбиелеп жатқандығын да атап айтуға болады.

Мемлекеттік тіл дәрежесіндегі бүгінгі қазақ тілінің қызметін күшейту ісінде тіл білімінде қолданбалы теорияларды дамытудың мәні зор, заман талабы екендігін күнделікті өмірден көріп отырмыз.

Тіл мамандарының ең талантты дегендерінің ішінде екінің бірі бара алмайтын бұл саладағы мүшелі той иесінің жаңадан жол салып, бастау көздердің бірі болуында екі-үш фактордың мәнін атау керек: бірі – бұл іске математик маманның кірісуі, екіншісі – математиктің тіл білімі сияқты гуманитарлық ғылым саласының негізгі қағидаларын, тармақтарын, теориялық тұғырлары мен практикалық қажеттіліктерін бірден жақсы игеріп (оқып), танып алған қабілет-дарыны, үшіншіден, қазақ тіл білімі саласын заман талабына сай дамытуға үлес қосу (дәлірек айтсақ, жол салу) сияқты ғылыми мақсат қоюы. Ал мұндай мұрат-мақсатты қоюға, сірә, әке жолын жалғастыру, «Жұбанов» деген ата атын жоғары ұстау сияқты жазылмаған адами заңның талабы да болған болар.

Әке жолы демекші, кейінгі жылдардағы ізденістерге қарағанда қазақ тілтанымының бүгінде қолға алынып, кеңінен зерттеу нысандарына айналып жатқан жаңа салаларының зерттеудегі жаңа әдіс-амалдарының алғашқы ұшқындары әкесі Құдайберген Жұбановтың еңбектерінен көрінетіндігі белгілі болып отыр. Әрине, оларда қазіргі терминдер, XX ғасырдың II жартысындағы ғылыми жаңалықтар орын алмағандығы да белгілі, ал негізгі идея ғылыми сұраныс-қажеттік дегендердің ұшқынын, идеясын проф. Құдайберген Жұбанов еңбектерінде айтарлықтай көзге түседі. Қазіргі кезде көбірек назар аударылып жүрген «валенттілік теориясы», «деривация құбылысы», «функционалдық грамматика», «лингвистикалық статистика» дегендердің алғашқы көрінісі проф. Құдайберген Жұбановтан табылатындығын Қ.Бектаев, Қ.Молдабеков, Т.Ысмайыловтардың, сондай-ақ А.Жұбанов пен басқа жас ғалымдардармен бірлесіп жазған мақалаларынан көреміз. Әсіресе бұл топта дәлелді де, айқын пікірлерді проф. Асқар Жұбанов айтып келеді.

Құдайберген Жұбанов балаларында, әсіресе Мүслима, Есет пен Асқар әкелерінің ұшығы көрінген идеяларын жүзеге асыру, әке бастап берген ғылыми тармақтарды әрі қарай дамыту, дұрыстығын дәлелдей түсу айрықша сезіледі. Бұл өзгелерге өнеге, кейінгі жастарға үлгі болатын іс-қарекеттер.

Табиғи талантын таныту, қолға алған жұмысын мұқият, адал орындау, әр іске үлкен жауапкершілікпен қарау, еңбегінің жемісін (жеңісін де) көрсете алу жағынан ағасы Есет сияқты, Асқар Жұбанов та әке алдында парызы мен қарызын өтеп отыр. Мұны Асқар сонау 1973 жылы кандидаттық диссертация қорғаған алғашқы адымында көрсеткен болатын. Әр алуан заңнан аттап қия баспайтын сол кезең түгіл, қазіргі кезде батылымыз бармайтын бір оқиға – сол қорғау үстінде бас сарапшы, атақты ғалым Р.Г.Пиотровскийдің бұл кандидаттық жұмысты өте жоғары бағалап, докторлық жұмысқа сай, бірден докторлық дәрежеге ұсынуға болар ма еді дегені болатын. Мұндай пікірді біз кейінгі-соңғы ешбір қорғауда естіген жоқпыз. Ал асықпай жүріп, үлкен жетістіктермен (тың тұжырымдар, теориялық-ұстанымдар, практикалық ұсыныстармен) 2002 жылы қорғаған докторлық диссертациясы да кезіміздегі ең таңдаулы, ең қажет ғылыми еңбек деп танылғанын көптеген рецензия-пікірлер, сол диссертацияның негізгі қағидаларына сүйеніп жазылып жатқан еңбектер көрсетіп отыр.

Асқар Жұбановтың докторлық дисертациясы тіл білімінің жаңа саласы мәтін лингвистикасы тұрғысынан жазылған. Қазақ тілінің келелі мәселелерінің біріне жататын мәтіннің мазмұндық құрылымының түзілуін абзацтық деңгейде қарастыруға және қазақ мәтінінің семантикалық көрінісін бейнелеу мен оның туындау мүмкіндіктерінің формалды принциптерін анықтауға арналған.

Бұл жұмыстың аса маңызды жақтары ҚР Жоғары аттестациялық комитетінің «Бюллетенінде» Президиум атынан арнайы аталып өткен болатын (№3, қыркүйек, 2003). Онда: зерттеу жұмысының нәтижесі қазақ тілінің қолдану аясын кеңейтуге өз ықпалын тигізетіні және оның қазіргі заманғы ақпараттық технологиялардың жаңа сұраныстарынан туындап отырғаны жайлы айтылған.

Асқар Жұбановтың еңбек авторларының бірі болып қатысуымен М.О.Әуезовтің «Абай жолы» романының жиілік сөздігі (1979), М.О.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері (Алматы, Түркістан, 1995), Қуманша-қазақша жиілік сөздік (Алматы, 1978) жарық көрді. Жиілік сөздіктер тек статистикалық мәлімет алу тұрғысынан ғана емес, лексикалық қазынаны құрайтын сөздердің қолданыс әлеуеті, ескі-жаңалығы, «сүйегі» (тегі), номинативтік (лексикалық, түпкі) мағынасынан өзге стильдік қызмет ету қарекеті сияқты тілтанымның өзге де салаларында пайдалануға болатын таптырмас қымбат материал болып саналады. Бұл тұрғыдан алғанда көрсетілген жиілік сөздіктер аса қажет құралдардың бірі.

Бұлардан басқа ғалымның негізгі теориялық-танымдық еңбектерін қайталап атасақ, олар мыналар: Квантитативная структура казахского текста (опыт лингвистического анализа на ЭВМ). – Алма-Ата: Наука, 1987. –147 с.; Основные принципы фомализации содержания казахского текста. –Алматы,

2002. 250 с.; Қолданбалы лингвистика: қазақ тілінің статистикасы (1-басылымы 2004, 2-басылымы 2007); Қолданбалы лингвистика: формалды модельдер (2006, 2007ж); Компьютерлік лингвистикаға кіріспе (2007), Қазақ сөзінің «**Тіл – қазына**» деректер базасы және оның теориялық негіздері – Алматы: «Арыс», 2009. – Б. 304. Ал әр алуан жинақтарда, журналдарда жарияланған жүзден аса мақалаларының да ғылыми құндылығы мен қажеттігін атауға болады.

БжҒМ А.Байтұрсынұлы атындағы Тіл білімі институтына аспирант болып келгеннен бері Асқар Жұбановтың бір ғылыми мекемеде қызмет етіп жүргеніне міне 45 жыл толып отыр екен. Аспиранттық кезеңін Минск қаласындағы шет тілдер Институты жанында өткізіп, кандидаттық диссертациясын қорғауынан сәтті өтіп, кіші ғылыми ғылыми қызметкерден бас ғылыми қызметкерге, бөлім меңгерушісі дәрежесіне дейін көтерілді.

Тіл білімі институтының табалдырығын аттағаннан бері А.Жұбанов қазақ тілінің қолданбалы саласының жаңа бағытына қатысты тақырыптарға ғылыми жетекшілік етіп келеді. Соңғы жылдары компьютерлік лингвистика бойынша қазақ тілінің «Тіл – Қазына» атты компьютерлік базасын іске қосу жолында аса нәтижелі жұмыстар атқаруда.

Әке жолы демекші, тағы бір таңғажайып құбылысты айтайық. Мен Құдайберген Жұбановтың өмірі мен ғылыми жолы туралы 1965 жылдан бері танысып, әр кез жазып, баяндамалар жасау барысында ғалымның феномендік тұлғасын көрдім де «сан қырлы, саңлақ ғалым» деп атадым. Сан қырлы дейтініміз – құдай оның бойына көп дарын дарытқан; алдымен, тілге деген үлкен таланты болған, сол таланттың арқасында 2-3 жылда орыс тілін меңгеріп алады, Ленинградта аспирантурадағы екі-үш жылда неміс, монғол, грузин, чуваш, коми, тіпті жапон тілдерін үйрене бастайды, ұзын-ырғасы екі жылдай медреседе оқып, араб, парсы, шағатай, түрік, татар, башқұрт тілдерін игеріп кетуі осы қабілеттің күші болар. ҚазПИ-де профессор болып жұмыс істей жүріп ұйғыр тілінде үйреніп, оның оқулығын жазып бермек болады, дунған тілінің маманын даярлауға кіріседі. Әкелерінің бойындағы осындай полиглоттық қабілет ұлдары Есет пен Асқардың тіл маманы, ал қызы Мүслиманың педагогика маманы болып қалыптасуларына ықпал етсе керек.

Әкенің тағы бір қасиеті – өнер иесі болғандығы: өнердің «көкесі» ақындық дарын, кейін дамытпаса да, бала кезінен бастап, жігіт шағында құрбы-құрдастарына әзіл-қалжыңдар жазып жүреді, ел аузынан өлең-жырларды жазып алып жинайды, өзі де өлең, әңгіме жазады т.т. Осы «қылық» Есет пен Мүслимада да бар екен. Мүслиманың әкесі туралы поэмасы көркемдігі жағынан да, идеясы мен мазмұны жағынан да ешбір әдеби туындыдан кем емес. Ал Есеттің ақындығына Тіл білімі институтының қызметкерлері тәнті болатын, ол қабырға газетін шығарған кезде екі-үш күнге дейін онда жарияланған Есеттің өлең шумақтары Институттың коридорын күлкімен жаңғыртып жататын.

Құдайбергеннің алғашқы ұстаздарының бірі Абдолла Беркінов: «Құдайберген сурет өнеріне өте шебер болды» депті. Темір, Жұрын

аудандарының қорымдарындағы құлпытастарда сақталған жазулардан Құдайберген Жұбановтың қолын қазірге дейін табуға болады. Әкелерінің суретке деген дарынын үлкен ұлы Ақырап иеленді.

Құдайберген Жұбанов ең алдымен ұстаз, тәрбиеші, педагогтік еңбегімен жақсы танымал болды. Тұңғышы Мүслима – осы жолды қалап алды. Профессордың бойында математика, экономика сияқты салаларға қызығушылық та болғанын зерттеушілер мен туған інісі Ахмет Жұбанов айтқан болатын. Асқар – математик, Қызғалдақ – экономист деген сияқты әке бойындағы дарын ұшқындарын балаларына аманат етіп кеткендей! Бұл – бір табиғат ұсынған ғажайып құбылыс, сол құбылыстың бір куәсі бүгінгі мерейтой иесі болып отырғанына алла тағалаға мың рақмет айтамыз.

Адам өзі өрекпіп көрсетпесе де, даурығып дәлелдемесе де, ең мықты ғалым оның бойында жатқан небір жақсы қасиеттері барлардан шығатын сияқты. Жұбановтар әулетінің қай өкіліне қарасақ та, үлгі тұтарлық кісілік ұстамдылық, талаптылық пен таланттылық, еңбекқорлық қасиеттерді көреміз. Өздерін жоғары орынға апарған да, біздерді сүйсіндірген де осы қасиеттер болар.

Өмірде кездесіп жататын өлім-қаза, ауру-сырқау сияқты ауыр жүктерді көтере білу де адамға үлкен сын. Осының бәрі нар көтерер жүк болса, сол жүкті де Асқар көтеріп келеді. Заман мен солақай саясат сынынан өте білген ғалым-азамат тағдыр сынына да арқасын төсеп отыр. «Өмірдің жақсы да, жаман да жақтарын басымнан кешірдім, бірақ оның қиын тұстарының бәрі де әкемнің нақақтан «халық жауы» аталуымен салыстырғанда түк те емес сияқты» – дейді Асқар бір толғанысында.

«Асылдың сыныағы, тұлпардың тұяғы», өзі де сом асыл – мықты ғалым, атаулы азамат, үлгілі ұстаз Асқар Құдайберген баласы, 75 жасың құтты болсын! Қазақ халқы мақтанышпен қастерлеп атайтын «Жұбановтар әулеті» деген қымбат атты жоғары ұстап, кейінгі ұрпақтарға жеткізе бер, ұзақ-ұзақ жасап, әлі де том-том еңбек жазып, топ-топ шәкірт дайындай бер деп апалық, үлкендік бата-тілегімізді білдіреміз.

Академик Рәбиға Сыздық

КІРІСПЕ*

Қазіргі кезде қоғамның барлық салаларында жаңа ақпараттық технологияларды қолдану өзекті мәселеге айналып отыр. Осыған байланысты Қазақстан Республикасының Президенті Н.Назарбаев өзінің «Қазақстан жолы» еңбегінде: «Алдағы онжылдықта жаңа технологиялық сипатты дамудың шешуші бағыттары био- және нанотехнологиялар, жасанды интеллект жүйелері, жаһандық ақпарат тораптары және жоғары жылдамдықтағы көлік жүйелері, қуат үнемдеуші технологиялар болып табылады. ....Өндірісті одан әрі интеллектендіру, үзіліссіз инновациялық үрдіске көшу жүзеге асырылады. Білімге негізделген жаңа үлгідегі қоғамға біртіндеп көшу аяқталады», – деп [1, 41-б.] айтқан болатын. Міне, осы аталғандарға қатысты қазақ тіл біліміне байланысты Кеңес одағы кезінде-ақ Қазақ ССР Ғылым академиясының Тіл білімі институтында 1969 жылы өткен «Түркі тілдерін статистикалық және ақпараттық зерттеу» атты Бүкілодақтық ғылыми жиыннан бастама алды деуге болады. Көп ұзамай, 1970 жылы математикалық статистика саласының маманы, белгілі математик-ғалым Қалдыбай Бектаевтың жетекшілік етуімен институтта «Тіл статистикасы және автоматтандыру» атты ғылыми топ құрылды. Осы кезеңнен бастап қазақ тілін статистикалық тұрғыдан зерттеу ісі жүйелі түрде қолға алынып, өз нәтижелерін бере бастады. Олардың алғашқысы ретінде 1973 жылы жарық көрген «Қазақ тексінің статистикасы» атты ғылыми жинақты атауға болады [2]. Сонымен қатар 1973-1974 жж. Одақтық «Тіл статистикасы» тобының ғылыми жетекшісі Раймонд Генрихович Пиотровский мен Қазақстандық профессор ғалым Қалдыбай Бектайұлы Бектаевтың «Математические методы в языкознании» атты екі бөлімнен тұратын, жоғары оқу орындарына арналған оқу құралының қазақ елінде жарық көруі де осы жаңа саланың жетістігі деуге болатындай [3].

Қазақ тіл білімінің жаңа саласы ретінде статистикалық лингвистиканың проблемалық мәселелері күн санап өсіп, олардың кейбір тұстары анықтала бастағаны мәлім. Міне, сол кездердегі (1978 жылы) «Тіл статистикасы және автоматтандыру» ғылыми-зерттеу тобына жетекшілік еткен Қ.Б.Бектаевтың «Статистико-информационная типология тюркского текста» атты докторлық монографиясы қазақ тіл білімінің статистикалық зерттеу саласына қосқан сүбелі еңбегі болды [4]. Аталған еңбек басқа да тілдерді зерттеу тәжірибесінде әр тілді жалпы типологиялық тұрғыдан бағалауда қолданылатын статистикалық критерийлердің үлес салмағын «өлшеуге» арналған іргелі ғылыми зерттеу ретінде саналды.

---

* «Кіріспе» ретінде **«Ахмет Байтұрсынұлы атындағы Тіл білімі институты. 50 жыл»** атты ғылыми жинақта жарияланған автордың «Компьютерлік лингвистика бөлімі» атты макаласын оқырман қауымына ұсынып отырмыз: ҚР Білім және ғылым министрлігі Ғылым комитеті. А.Байтұрсынұлы атындағы Тіл білімі институты. Алматы «Шапағат-Нұр» 2011, 217-237-бб. – *Редколлегия.*

Сол сияқты, 1978 жылы отандық түркітану ғылымы үшін ортағасырлық жазба ескерткіші «Кодекс Куманикус» мәтіні тұңғыш рет ЭЕМ жадына енгізіліп, ол бойынша ЭЕМ көмегімен «Куманша-қазақша жиілік сөздігінің» алынуы қазақ тіл білімі үшін аса бағалы зерттеу жұмысы болды [5]. Ал 1979 жылы «Ғылым» баспасынан жарық көрген М.О.Әуезовтің «Абай жолы» романының жиілік сөздігі» атты ұжымдық лексикографиялық еңбек те қазақ ғалымдарының статистикалық лингвистика саласына қосқан айтулы еңбегінің нәтижесі екендігі даусыз [6].

Қазақ елінің тәуелсіздігі жарияланған 1991 жылы А.Байтұрсынұлы атындағы Тіл білімі институтында «Түсіндірме сөздік» бөлімі мен «Тіл статистикасы және автоматтандыру» ғылыми тобы біріктіріліп, «Қолданбалы лингвистика» бөлімі ретінде жаңаша аталып, қайтадан құрылды. Міне, сол кезден бері жаңа бөлімнің зерттеу мақсаттары мен міндеттері ұлғая түсті. Қазақ қолданбалы лингвистика бөлімі қызметкерлері және ғылыми меңгерушісі (осы макала авторы) өз ұстазы проф. Қ.Б.Бектаевтың ғылыми жолын жалғастыру мақсатымен, қазіргі қазақ тілінің жаңа саласы қолданбалы тіл білімінің теориялық және практикалық негізін қалау ісіне айтарлықтай үлес қосты деуге негіз бар. Мәселен, өткен ғасырдың 90-шы жылдары 20 томнан тұратын М.Әуезовтің шығармалар жинағы ЭЕМ жадына енгізіліп, соның нәтижесінде 1995 жылы «М.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері» атты жиілік сөздіктер жүйесі жарық көргенін айтуға болады [7]. Міне, осындай жиілік сөздіктердің негізінде ұлы жазушының тіл байлығы мен стильдік ерекшелігі жайында бірнеше ғылыми зерттеулер жүргізіліп, диссертациялар қорғалды.

1991-1993 жылдар аралығында бөлімде «Қазақ тілінің қолданбалы саласындағы проблемалар» атты тақырыпты орындауда филол.ғ.д., профессор Ә.Болғанбаев, А.Жұбанов, Ш.Дәулетқұлов, С.Бизақов сияқты ғалымдар өнімді еңбек етті.

1998 жылы аталмыш бөлім Институттың құрылымдық өзгерістеріне байланысты (кейбір бөлімдердің қысқаруына байланысты) енді «Компьютерлік лингвистика» ғылыми тобы деп аталып, қаржыландыру жағынан фонетика бөлімінің құрамына енді. Атауы мен мәртебесі өзгергенімен, зерттеу бағыт-бағдары айқын ғылыми топ өз қызметін әрі қарай тиімді жалғастырды деуге негіз бар. Себебі 2002 жылы ғылыми-зерттеу тақырыптардың жетекшісінің (мақала авторы) «Основные принципы формализации содержания казахского текста» атты докторлық монографиясы қазақ қолданбалы лингвистикасының болашақ дамуына қосылған елеулі үлес ретінде саналды [8].

Аталған еңбекте қазақ мәтінін формалды әдіспен зерттеу тәсілі қазақ лингвистикасында тұңғыш рет қолданылған болатын. Мұндай зерттеу тәсілінің қажеттігі заман сұранысынан туындайды, яғни тәуелсіз Қазақстандағы қазақ тілінің қызметіне, республикадағы коммуникативтік аяға жаңаша көзқарасқа, компьютерлік және телекоммуникациялық технологияның кең тарауына және т.б. байланысты қазақ тілін жаңа қолданбалы әдіс-тәсілдермен зерттеу қажеттігінен туындады.

Еңбекте қазақ тіл білімінің өзекті мәселелеріне жататын мәтін және оның формалдық, мазмұндық құрылымына қатысты құрастырушы элементтер арнайы зерттелген. Зерттеудің негізгі мақсаты – қазақ мәтінінің семантикалық бейнесі мен оны туындататын негізгі формалды принциптерді анықтау, яғни қазақ мәтінін құрастыру мен туындатудың формалды жақтарын айқындау. Автордың қазақ мәтінінің статикалық және динамикалық мазмұнын бейнелеу тәсілін жоғары оқу орындарының оқулықтар мен оқу құралдарын құрастыруда қолдануға әбден болады.

Монографиялық зерттеуде қолданбалы тіл білімінің жаңа саласы – «Мәтін лингвистикасына» сәйкес мәтін бірліктерінің таңбалық сипаты, абзацтардың мазмұнға қатысты ерекшеліктері мен қызметі анықталған. Қазіргі тіл зерттеу тәжірибесіндегі модельдеу әдісіне сәйкес «табиғи тіл – әлем моделі» қатынасы тілдің синтаксистік және семантикалық, статистикалық модельдерін құрастыру арқылы дәйектелген. Сонымен бірге қолданбалы лингвистиканың жаңа салалары: жасанды интеллект, машиналық аударма, ақпарат іздестіру жүйелері, қазақ тілінің компьютерлік қорын құрастыру жайттары және т.б. қарастырылған.

Алынған нәтижелердің негізінде қазақ лингвистикасында жаңа қолданбалы бағыт негіздері қалыптасады. Нәтижесінде қазақ тілінің қолданбалы бағыты корпустық лингвистика, компьютерлік, квантитативтік лингвистика сияқты жаңа салалармен толығып, қолдану аясы кеңейеді, жаңа әдіс-тәсілдерді молынан қолдануға мүмкіндік туатыны анықталды.

Автор ұсынған қазақ мәтінін формалды пішінге келтіруде қолданыс табатын жасанды семантика-синтаксистік тіл (СЕМСИНТ) республиканың әлеуметтік-экономикалық дамуына қатысты көптеген өзекті мәселелердің шешімін табуына оң әсер ететіндігі белгілі болды.

Сонымен, 1970-2000 жылдар аралығындағы ғылыми ізденістер, негізінен, жеке автор шығармаларын, функционалдық стильдерді, сондай-ақ толық әдеби тіл қамтылған мәтіндерді статистикалық әдіс-тәсілдермен зерттеу тәжірибесінде әртүрлі жиілік сөздіктер жасаудың өзекті (актуалды) мәселе екендігін көрсетті. Зерттеуде қолданбалы лингвистика саласының әдіс-тәсілдері қамтылып, жиілік сөздік мәселесі, олардың автоматты жасалу жолдары көрсетілген. Қазақ мәтініндегі негізгі сөз таптары үлестірілуінің ықтималды-статистикалық моделін құру мәселесі зерттелген.

Зерттеу нәтижелері негізінде аса маңызды және күрделі мәселелерді: тілді лексика-грамматикалық нормалауға, түркі тілдерін өзара салыстырмалы-типологиялық зерттеуге және олардың компьютерлік қорын құруға зор мүмкіндік жасайтындығы дәлелденді.

Компьютерлік лингвистика ғылыми тобының негізгі зерттеу бағыты Тәуелсіз еліміздің мемлекеттік тілін қоғамның барлық саласында ғылым тілі ретінде қолдануға бағытталған «компьютерлендіру» ісіне арналады. Қазақ тілінің көптеген кезек күттірмейтін мәселелері, мәселен, лексикография саласы бойынша компьютерлік қор (база) жасау, қазақ мәтіндері корпустарының компьютерлік базасын құрастыру және қолданбалы лингвистика саласының басқа да мәселелері «Компьютерлік лингвистика»

ғылыми тобының 2000 жылдардан бастап жүргізген ғылыми ізденістерінен көрініс тапты.

Мәтінді туындату заңдылықтарын білудің практикалық маңызды жағы – орта және жоғары білім беру жүйесін жаңаша ұйымдастыру мәселесі мен білім берудегі жекелік және топтық әдістер арасындағы қайшылықтарды шешуге қатысты мүмкіндіктер. Міне, оқыту процесін осы сияқты мәселелерге қатысты ұйымдастыру жұмысына жаңаша көзқарас қажет болды. Аталған мәселелерді шешу үшін әр түрлі мәтіндердің құрамдық сипатын, олардың туындау заңдылықтарын білу және мәтін ішіндегі негізгі және көмекші тілдік элементтерді айыра білу де қажет болады. Ал бұл, бір жағынан, оқушылардың қажетті деген оқу материалдарын игеру ісінде, екіншіден, мамандардың жаңа оқулықтар мен оқу құралдарын құрастыру кезінде де басшылыққа алатын ұстаным еді.

Жоғарыда аталған мәселелер барлық түркі тілдерін, соның ішінде қазақ тілін зерттеу ісіне де тікелей қатысы бар екені мәлім. Әсіресе, қазақ тілінің мемлекеттік дәрежесін алуы, оның халық шаруашылығын басқаруда, оқыту мен тәрбиелік жұмыстарында және қазақ халқының мәдени өмірі мен егемен елімізді қазақ тіліндегі ақпаратпен қамтамасыз ету мәселелерінде аса маңызды болмақ.

Егеменді еліміздегі қай ғылым саласын алмасақ та, оның қарқындап дамуы үшін ең алдымен халық шаруашылығын қазақ тілінде басқаратын және оқыту процесіне қажетті сапалы (әрі жедел) түрде жұмыс атқаратын автоматтандырылған жүйелерді құру мәселелері өмір талабына сай әлемдік дәрежеге көтерілуі тиіс деп санаймыз.

Әрине, мұндай мақсатқа қол жету үшін қазақ тілінің дәстүрлі грамматикаларды формалдылық тұрғысынан қайта қарастырып шығу қажет. Себебі қазақ тілінің табиғатында, басқа тілдер типтерінде де көптеп кездесетін, күрделілік пен бірыңғайлыққа жатпайтын қасиеттер және тілдік бірліктерді формалды жолмен бөліктеудің, таңбалаудың, мән-мағынасын ашудың және т.б. жайттарды анықтау мәселелерінде қиындықтар көптеп кездеседі.

Бұл айтылғандар тілдік бірліктердің тұрпаты (формасы) мен мән-мағынасы арасындағы қатынастан, яғни сөзтұлға мен оның мағынасы, сөйлем тұрпаты мен оның мән-мағынасы арасынан және мәтін мен оның мазмұны арақатынастары сипаттарынан байқалады. Қазақ тілінің мұндай бірізділікке жатпайтын тұстары сөздердің не сөз тіркестерінің мағыналы құрылым жақтарында да баршылық. Міне, мұндай бірізді шешім таппайтын сипаттамалар қазақ мәтінін автоматты түрде талдау мен жинақтау жүйелеріне және білім беру саласындағы жоғарыда аталған автоматты жүйелердің дұрыс жұмыс атқаруына барынша кедергі келтіретіні сөзсіз.

Қазақ тілінің грамматикалық құрылымын дәстүрлі әдіспен зерттеу арқылы қол жеткен теориялық және практикалық нәтижелер мәтін лингвистикасы саласындағы жаңа ғылыми ізденістерге негіз болатыны белгілі. Сондықтан мәтін мазмұнын компьютер арқылы түсінуге мүмкіндік тудыратын модель жасау үшін мәтін құрамындағы негізгі тілдік бірліктердің

(морфема, сөз, сөз тіркес, сөйлем және т.б.) мағыналарын формалды түрде анықтауға қатысты зерттеулер А.Байтұрсынұлы атындағы Тіл білімі институтында іргелі және қолданбалы тақырыптар бойынша жүйелі түрде жүргізілуде.

Қолданбалы лингвистика бөлімі қазақ тілін автоматтандыру ісінің жаңа сатысына көтеріліп, 1991-1993 жылдары «Қазақ тілінің «ҚАЗЫНА» атты автоматтандырылған картотекалық базасын жасау» атты тақырыпты қолға алғандығы қазақ тіл білімінде ғана емес, түркі тіл біліміндегі орасан зор жаңалық болды деп айта аламыз. Осы жылдары бөлім қызметкерлері Тіл білімі институтында жасалып жатқан «Қазақша-орысша сөздікті» дайындауға да барынша атсалысты.

Сонымен көп жылдар бойы жинақталып-теріліп, қазақ тілінің бірден-бір «тіл байлығы» деуге болатындай, институттың «сөз қоржынында» сақталған қазақ тілінің картотекалық қорын сан мен сапа жағынан сұрыптап, компьютер жадына түсіріп, автоматты түрде пайдалану ісінің қажеттігі тек бүгінгі таңға ғана емес, оның болашағының маңыздылығында. Қағаз түріндегі кәртішкі-мәнмәтін қордың жалпы көлемі, орташа есеппен алғанда, 5 млн. шамасындай. Әрине, осындай үлкен көлемдегі мәнмәтін қорын автоматты жүйеге ауыстырмайынша, оны тілші қауымының ойдағыдай пайдалану мүмкіндігі және үнемі жаңартылып отыруы оңайлыққа түспейіні мәлім.

Міне, осындай себептерден өткен ғасырдың 90-шы жылдарынан бастап қолға алынған қазақ тілінің «Қазына» (қазіргі «Тіл – қазына») атты автоматтандырылған картотекалық қордың маңызы айтарлықтай зор. 1991-1993 жылдар аралығында «Қолданбалы лингвистика» бөлімінің қызметкерлері «Қазына» қорына 900 мың кәртішкі-мәнмәтін ішінен 70 мыңдайын сұрыптап алып, оларды компьютер жадына енгізген болатын. Қазіргі кезде бастапқы әріптері «А» әрпінен «С» әрпіне дейінгі сөздерге қатысты мәнмәтіндер «Тіл – қазына» компьютерлік базаға енгізіліп, тұтынушы-лексикографтар үшін қажетті деген мәліметті іздеп табу жұмыстары жүзеге асырылып отыр. Ал «Т» әрпінен «Я» әрпіне дейінгі кәртішкі-мәнмәтіндерді база енгізу әрекеті қазіргі кезде өз жалғасын табуда.

«Тіл – қазына» атты қазақ тілінің кәртішкі-мәнмәтіндік компьютерлік базаның негізгі мүмкіндіктерін қысқаша сөз ететін болсақ, олар:

- қажетті деген сөзге немесе сөзформаға қатысты кәртішкі-мәнмәтіндерді лезде тауып алып, олармен компьютердің экраны арқылы немесе қағаз бетіне басып шығарып та танысу мүмкіндігі;

- әрбір әріпке байланысты реестрлік сөздік түзілгендіктен, қажет деген сөздер мен оларға қатысты кәртішкі-мәнмәтіндерді лезде іздеп тауып алу мүмкіндігі. Мұндай әрекетті кәртішкі-мәнмәтін ішіндегі сөзқолданыстар бойынша да жүзеге асыру алдын ала қарастырылған;

- қажетті деген сөзге немесе сөзформаға (олардың бастапқы, ортаңғы, соңғы бөліктеріне) байланысты кәртішкі-мәнмәтіндерді авторына, шығармасына, баспа аттары мен шыққан жылдарына қатысты іздеп табу мүмкіндігі;

- «Тіл – қазына» компьютерлік базадағы кәртішкі-мәнмәтіндерді өңдеуге, өшіруге немесе жаңадан қосуға да мол мүмкіндік ескерілген.

Қазақ тілінің «Тіл – қазына» атты компьютерлік базасы картотекалық қордың барлық міндетін атқарумен бірге, ол енді сөздіктер мен мәтіндерді бейнелеу қорына да айналып отыр.

А.Байтұрсынұлы атындағы Тіл білімі институтының іргелі (2006-2008жж.) және қолданбалы (2008-2009жж.) зерттеу бағдарламалары бойынша «Компьютерлік лингвистика» ғылыми тобының атқарып жатқан ғылыми жұмыстарының нәтижелері осының айғағы. Қазақ лексикасының түсіндірме сөздіктері негізіндегі компьютерлік базасын іске қосу мақсатына сай зерттеу жұмысының нысаны Тіл білімі институтында құрастырылып, 1968 жылы жарық көрген «Абай тілінің сөздігі» мәтіні (6 мың сөз, 1 мыңдай сөз тіркес), 1999 жылы баспадан шыққан «Қазақ тілінің сөздігі» (50 мың сөз) және жаңадан құрастырылып жатқан 15 томдық «Қазақ әдеби тілінің сөздігінің» мәтіні. Сол сияқты «Ұлттық идея – Қазақстанның даму негізі» бағдарламасы бойынша: «Қазақ тілінің терминологиялық қорының компьютерлік базасын жасау және оны терминдік әлеуеті жоғары дереккөздермен толықтыру» атты ғылыми-зерттеу жұмысының негізгі нысаны салалық екітілді (қазақша-орысша, орысша-қазақша) терминологиялық сөздіктер мен қазақ тіліндегі салалық терминологиялық түсіндіме сөздіктер материалдары.

Мұндай компьютерлік қор жасаудың негізгі мақсаты – зерттеуші-тілшінің жұмыс орнын толығымен автоматтандыру. Осының арқасында зерттеушінің уақытын барынша үнемдеу, яғни қажетті тілдік ақпаратты іздеу жұмысы мен тілдік бірлікті реттеу және т.б. жеңілдету тұстары қарастырылған.

Сонымен бірге тілдік процестерді автоматтандыру мәселесі тоқырап қалмай, үнемі жетілдіру жағынан өзінің жалғасын тауып тұруы қажет. Егер ана тілімізге күн сайын жаңа сөздер, жаңа қолданыстар мен жаңа терминдердің еніп отыратынын ескерсек, мұндай зерттеу – күн тәртібінен түспейтін әрі ұтымды, әрі заман сұранысына сай зерттеу деп білеміз.

Автоматты сөздік базасын құрастыру мәселесі қазақ лексикографиясы үшін толығымен шешілген жағдайда, сөздікті екі түрлі тұрпатта – кітап түрінде және сыртқы ақпарат сақтау таспаларына жазып алынған электрондық пішінде сақтауға мүмкіндік туып отыр. Нәтижесінде сөздіктердің қолдану аясы кеңейеді және лексикография саласының қызметі де ұлғая түседі деуге болады.

«Тәуелсіз Қазақстан жағдайында қазақ тілінің ұлттық және мемлекеттік тіл ретінде негізгі даму үрдістері мен бағыттары» атты іргелі зерттеу бағдарламасы бойынша «Компьютерлік лингвистика» ғылыми тобы 2006-2008 жылдары жүргізген **«Қазақ тілінің компьютерлік базасы: лексикалық қордың ашық жүйесі»** атты ғылым-зерттеу жұмысы 2003-2005 жылдары жүргізілген зерттеулердің қисынды жалғасы болып саналады.

Еліміздің егемендік алуына байланысты бүгінгі таңда ана тіліміздің өз мәртебесіне сай қоғамның барлық саласында орнын алу барысы жан-жақты жүріп жатыр. Осыған орай, толық сипаттағы болашақ қазақ тілінің «Тіл –

қазына» деректер базасының ішінде неше түрлі мәтіндер, сөзтұлғалардың статистикасы, сөзнұсқағыштардың толық тізімдері, сөздерге телулі цитаталар, жиілікті және кері-әліпбилі сөздіктер көрініс табатыны сөзсіз. Аталған тақырып өзегіне қатысты барлық еңбектер, ғылыми зерттеу жұмыстары бірізге, бір жүйеге түсіп, зор жетістіктерге жетеді деп білеміз.

Кез келген тілдің сөздік құрамы жүздеп-мыңдап саналады. Сондықтан қажетті сөзді тез іздеп табу үшін лексиканың жүйелік сипатын ескеретін стратегияны туындатып алған жөн болады. Соның негізінде қажетті деген сөзді барлық сөздік құрамының ішінен іздемей, тек оның кішігірім бөлігінен, мысалы, семантикалық өріс ішінен немесе лексика-семантикалық топ ішінен және т.б. іздеу мүмкіндігі туындайды. Оқу лексикографиясының теориясы, ең ұтымды жағдайда, тіл жүйесінің лексикалық аспекті бойынша, аспектілік сөздіктер жүйесін құруды қалайды. Сол арқылы іздеудің табиғи стратегиясы модельденіп, қажетті сөзді іздеп-табу мезгілі қысқарады.

Бұл мәселе сөздіктің көлеміне, оларды ұйымдастыруға, сөздіктегі әрбір лексикалық бірлікті сақтайтын ақпарат көлеміне, әрқилы таңбалар жүйесіне, сілтемелер жүйесін құру жұмыстарына да қатысы бар. Яғни бұл мәселе кешенді сипатқа ие болу керек деген сөз.

Қазіргі кезде халықаралық қатынас нығайып, көптеген дербес ұлттық мемлекеттердің пайда болуы және бұрынғы отарлықта қалған халықтардың мәдениетінің көтерілу сипаты, қайтадан лексикографияның тіларалық қарым-қатынасқа және тілдерді оқытуға байланысты қызметтерін бірінші орынға ауыстыруына жағдай жасауда. Әлемдік өндірісті интеграциялау мәселесінде әртүрлі түсініктерді бірізге түсіру ісі де аса қажет болуда. Сондықтан егемен Қазақстанның халықаралық деңгейге шығуына және бірқатар жетекші елдермен қазіргі таңдағы ғылым мен технологияның әртүрлі салаларында ғылыми және өндірістік ақпараттармен алмасудың едәуір ұлғаюына байланысты неғұрлым кең ауқымды қамтитын көптілді сөздіктер шығару қажеттілігі де туындауда. Жақын болашақтағы лексикографиялық зерттеулердің негізінде оқу сөздіктерінің әр ұлттық тілдерге, соның ішінде қазақ тіліне қатысты да практикалық маңызы зор жүйесі құрылуы мүмкін.

Сонымен 2006-2008 жылдары зерттелген **«Қазақ тілінің компьютерлік базасы: лексикалық қордың ашық жүйесі»** атты тақырып бойынша жүзеге асырылатын жаңа пішіндегі және жаңа мүмкіндіктерге ие болып отырған «Тіл – қазына» атты қазақ тілінің картотекалық және түсіндірме сөздіктері бойынша әрекет ететін компьютерлік базасының мүмкіндіктері мол және оны әлі де болса арттыра түсу қажет деп санаймыз.

Аталған тақырып бойынша орындалған ғылыми-зерттеу жұмысының алға қойған мақсаты мен міндеті ретінде қазақ лексикасының түсіндірме сөздіктері мен картотекалық қоры негізіндегі **«Тіл – қазына»** атты деректер базасын әлемдік талаптарға сай деңгейде құрастыру болатын. Міне, осыған байланысты **«Тіл – қазына»** деректер базасының алгоритмдік құрылымның теориялық негіздері зерттеліп, талапқа сай келетін база құрастырудың жасанды бағдарламалық тілі таңдалып алынды.

Қазақ лексикографиясын әлемдік дәрежедегі жаңа ақпараттық технологияға өтуіне байланысты «Автоматты сөздіктердің» атқаратын міндеті мен құрылымына қатысты, қазақ мәтінін автоматты өңдеуге қатысты, автоматтанған қазақ тіліндегі ақпаратты іздестіру жүйесінің мақсаты мен міндеттерін айқындауға қатысты және қазақ сөзі деректерінің тілдік бірліктердің **«Тіл – қазына»** атты базасын құруға және т.б. мәселелерге қатысты ғылыми әдебиеттермен таза лингвистикалық тұрғыда және аталған автоматты жүйелерді құрудың таза инженерлік тұстарымен, яғни компьютерге бағдарлама жасайтын маманға қажетті мәселелерінің теориялық және тәжірибелік тұстары толық зерттелді. Нәтижесінде 2009 жылы тақырып жетекшісі А.Қ. Жұбановтың «Қазақ сөзінің «Тіл – қазына» деректер базасы және оның теориялық негіздері» атты ғылыми монографиясы жарық көрді [9].

Сонымен, аталған зерттеутақырыбы бойынша қазақ тілінің лексикалық қорының түсіндірме сөздіктері негізінде компьютерлік база құру мәселесі қарастырылды. Компьютерлік базаның лексикалық нысана ретінде: 1) қазақ тілінің картотекалық қоры; 2) «Абай тілінің сөздігі»; 1 томдық «Қазақ тілінің сөздігі»; 3) 15 томдық – «Қазақ әдеби тілінің сөздігі» материалдары толық қамтылды. Қазақ тілінің лексикалық қорын ашық жүйе ретінде қарастырып, компьютерлік базаның мүмкіндіктерін: өзгерту, қосу, іздеу, адам мен компьютер арасындағы коммуникация процестерін барынша жетілдіру мәселелері қарастырылады. Зерттеу тақырыбының негізгі мақсаты – лексикография мен лексикологияға қатысты тілдік мәселелерді автоматтандыру, яғни қолдағы бар мәліметтермен тез және сапалы түрде танысу, қажетті деген әрекеттерді барынша жеңілдету. Сонымен бірге, қазақ тілінің лексикалық байлығын, компьютерлік база ретінде бүкіләлемдік «интернет» жүйесіне қосуға мүмкіндік бар. Міне, осы аталғандар қазақ сөзінің «Тіл – қазына» атты деректер базасын жүзеге асыру мақсатындағы ең маңызды мәселелер қатарына жатады.

2009-2011 жылдар «Компьютерлік лингвистика» ғылми тобы қызметкерлері **«Мәдени құндылықтар ретіндегі қазақ тіліндегі мәтіндер корпусы және сөздіктердің «Тіл – қазына» атты ұлттық компьютерлік қоры»** атты ғылыми-зерттеу тақырыбын өз зерттеу нысаны ретінде алды.

Аталған тақырыптың мақсат-міндеттеріне, нәтижелері мен жаңалықтарын сөз етерден бұрын, біз әлемдік дәрежедегі компьютерлік мәтіндер корпусын жасаудың теориялық және практикалық жақтарының лингвистикалық және инженерлік деңгейлеріне қатысты мәселелерге және корпустық лингвистиканың қазіргі тілтаным парадигмасында алатын орнын анықтау үшін оның қысқаша тарихына тоқталуды жөн көріп отырмыз.

«Корпустық лингвистика» Н.Хомскийдің көзқарасына қарама-қарсы бағытта пайда болды деуге әбден болады. Себебі, оның оппозициялық «құзырет/түсінігі» («компетенция/представление») бойынша тіл жайындағы деректер тіл иесінің құзыретін құрайды да, сондықтан да әртүрлі аса көлемді тілдік нысанды зерттеудің орнына сөйлеушінің тілдік интуициясына (ішкі түйсігіне) сүйенгенін жеткілікті деп санайды. Корпустық лингвистика мен

оған қатысты пікірталастар, айтарлықтай дәрежеде, теоретикалық және эмпирикалық білім аяларының ең ұтымды арақатынастарын іздестіру жолын бейнелейді деуге болады.

Корпустық лингвистика 1963 ж. АҚШ-та өмірге келген Браун корпусынан (The Brown Standard Corpus of American English) бастама алады. Бұл корпустың көлемі 1 млн. сөзқолданыстан тұрады және оның құрамында әрбіреуі 2 мың сөзқолданысқа тең 500 мәтін бар. Ғалымдар У.Н.Фрэнсис пен Г.Кучер бұл корпусты құрастыру үшін мынадай принциптер мен критерийлерді ұстаным етті:

1. Мәтін иесі ретінде американдық авторлармен ағылшын тілінің американдық нұсқасын қолданатын авторлар шығармаларын алу;
2. Мәтіндерді синхрондау (мезгілдік үйлестіру) үшін АҚШ-та 1961 жылы алғаш рет баспадан шыққан еңбектерді алу;
3. Жанрлардың арақатынасын таңдау жұмыстары үшін ерекше ықтималдықтар рәсімін қолдану: барлығы 15 жанр, 9 – «ақпараттық» проза, 6 – көркем проза; әрбір жанрдан 6-дан 80-ге дейін элементар таңдама мәтін алу;
4. Мәтіндерді таңдауға қойылатын талаптарды айқындау мен тұжырымдау: әр мәтіннің көлемі статистикалық тұрғыда, оның стильдік ерекшеліктерін ақиқат дәрежеде бейнелеуін ескеру.
5. Жанрлардың құрамы мен арақатынасы, олардың стильдік ерекшеліктері мен қатынастық салмағы барабарлық деңгейде көрініс табуы қажет [9].

Сонымен, Браундық корпус осыған ұқсас корпустар құруға қатысты зерттеулердің кеңінен тараған (атышулы) нысаны мен стандартына айналды. Ғалымдар арасында көптеген лингвистикалық зерттеулерді сапалы жүргізу тек ауқымды тілдік материалдар негізінде ғана жүзеге асатындығын ұғыну байқала бастады. Осы айтылғандардың барлығы мәтіндерді корпус түрінде ұйымдастыру ережелері мен оларға талдау жүргізу әдіснамасын зерттейтін бағыттың пайда болуына себепші болып, ғалымдарды одан әрі ынталандыра түсті. Сонымен корпустық лингвистика осы әдіс-тәсілдердің әдіснамасы ретінде туындап отыр деуге әбден болады.

Корпустық лингвистиканың ағылшын тілі материалында туындауы мен қалыптасуын ғалымдар АҚШ-та компьютерлік техника мен XX ғасырдың 60-80 жылдары британ лингвистикасындағы интеллектілік климаттың белсенді даму жағдаятымен түсіндіреді. Осы кездерде компьютерленген мәтіндік корпустарға лингвистикалық талдау жүргізу тілдік зерттеулердің ең көп бөлігін қамти бастағаны мәлім. Ғылымды институттендіру нәтижесі 2001 жылы Бирменгем университетінде корпустық лингвистика кафедрасын ашуға және International Journal of Corpus Linguistics журналын баспадан шығарып тұруға мүмкіндік туғызады. Бірнеше тілдердің материалдарында құрастырылған түрі мен қызметі жағынан айырым табатын корпустар және солардың негізінде құрастырылған неше түрлі сөздіктер мен грамматикалар өмірге келді.

Технологиялық құралдар ретінде компьютерлік технологияны қолданатын тіл білімінің басқа салаларымен қатар, корпустық лингвистика тілдік материалды жинау мен сақтаудың дәстүрлі әдіс-тәсілдеріне – жазба мәтінді қол күшімен өңдеу, лұғаттық картотекаларды жинау мен сақтау және т.б. түрлеріне тікелей әсер етті. Дәстүрлі тәсілдердің кемшілік жақтарына жататындар: аса көп уақыт шығыны, зерттеу нысанының көлем жағынан шектелуі, жаңартудың күрделігі, статистикалық өңдеудің қолайсыздығы, қашықтық жағдайындағы ақпарат алмасу мүмкіндігінің болмауы. Тілдік зерттеулерге қажетті материалдарды жинау мен сақтауға қатысты компьютерлік технологияны пайдалану жаңа тәсілдің артықшылығын толығымен пайдалануға мүмкіндік тудырды. Бұл жағдайда тілдік материалға «қол жеткізу» жедел түрде іске асады, интернет желісін пайдалану ақпарат алмасу қашықтығын жояды, ақпарат көлемі талай мәрте өседі, деректерді сақтайтын және өңдейтін компьютерлік бағдарламалар арқылы қажет деген ақпаратпен танысу мен оны өңдеу әрекеттері барынша тез жүзеге асады. Сонымен бірге бұл саладағы технологиялық олқылықтар ақпарат көлемінің шексіз ұлғаюының әсерінен кейбір тіл қызметінің ақиқаттық көрінісінен ауытқуларына әкеп соқтыруы да мүмкін екендігі байқалды [10].

Корпустық лингвистиканың негізгі міндеттері ретінде мына баптарды жатқызуға болады:

1. Аталған бағыттың теориялық негіздемесін жете зерттеу;
2. Корпустар түрлерін құрастыру мен қолдану тәжірибесін талдау;
3. Корпусқа қойылатын жалпы талаптарды тұжырымдау;
4. Әртүрлі зерттеулер мен оқыту міндеттерін атқаратын корпустар құрастыру;
5. Тіл білімінің әртүрлі салаларында мәтіндер корпустарын қолданудың тиімді тәсілдерін қалыптастыру. Бұл міндет зерттеуші қауымына практикалық тұрғыда аса құнды болып саналады.

Корпустық лингвистика тіл білімінің жеке саласы ретінде өзімен іргелес жатқан тіл ғылымы пәндерімен жанасып жатады, яғни математикалық лингвистика, дискурстік анализ және лексикография салаларымен жақын жатады. Корпустық лингвистиканың басқа тіл ғылымы пәндерімен қарым-қатынаста болу ерекшелігі, бір жағына алғанда, мәтіндер корпусының корпустық лингвистика қызметінің нәтижесі ретінде болса, ал екіншіден – лингвистикалық пәндердің басқа түрлеріне де бастапқы эмпирикалық материал болу мүмкіндігінде. Міне, дәл осы жағдай корпустық лингвистиканың фонетика, лексикология, грамматика және стилистика салаларымен тығыз байланыста болатындығына негіз бола алады.

Корпустық лингвистика әдістері ішінен келесі бөліктерді жеке қарастыруға болады:

1) филологиялық әдістер;
2) теоретика-лингвистикалық әдістер;
3) математикалық (статистикалық) әдістер;
4) ақпараттық технологиялардың әдістері.

Бірінші топ әдістеріне жанрға, мәтін авторларына, жанрлар мәтіндерінің үлесіне және т.б. мәселелерге қатысты әдістер жатады. Екінші топта – мәтін таңдаудың және корпустағы лингвистикалық ақпараттың берілуінің жалпы және жеке лингвистикалық мәселелерімен байланысты әдіс-тәсілдер. Үшінші топты таңдама мәтіндердің санына, оның репрезентативтілігін (тұлғалылығын) қамтамасыз ету аспектілеріне қатысты математикалық статистиканың әдістері құрайды. Төртінші топқа корпус деректерінің компьютерлік берілісі (көрінісі) мен оны өңдеу әрекеттерін қамтамасыз ететін ақпараттану әдістері жатады.

Корпустық лингвистиканы жаңа білім аясы деп түсінуге байланысты «әдіс» ұғымының орны мен рөлі айтарлықтай өзгеріске ұшырауы мүмкін. Мәселен, корпустық лингвистиканы компьютерлік лингвистиканың бір бөлігі деп қарастыратын болсақ, онда төртінші топқа қатысты ақпараттану әдістері басымшылық етер еді. Ал В.П.Захаровтың жоғары оқу орындарында «Корпустық лингвистика» пәнін оқыту курсында мәтінге лингвистикалық шартты белгілер енгізуге, корпустарды құру технологиясына, корпустан қажетті ақпаратты іздеу әдістеріне (конкордансерілерге, корпустық менеджерлерге, сұрату тілдеріне) көбірек көңіл бөлінген [11; 14].

Мәтіндер корпусын қолдану аясы неше түрлі болғандықтан, олар лингвистикалық, филологиялық және педагогикалық салаларға топтастырылуы мүмкін. Ондай топтарға жататын пәндер: лексикография, грамматика, мәтін лингвистикасы, типология, контрастивтік лингвистика, аударма жасаудың теориясы мен практикасы, әдебиеттану ғылымы, лингводидактика. Қазіргі кезде корпустық лингвистиканың қолдану аясы бұрынғыдан да кеңейе түсті.

Корпустық лингвистиканың негізгі түсінігін, оның бейнелейтін аясына қарай, корпустағы сақталатын ақпарат бірлігі тұрғысынан, оны әртүрлі негізде топтастыру мен құрастыру және т.б. принциптерді ұстаным етуіне қарай әртүрлі ұғынуға болады.

А.П.Барановтың тұжырымдауынша, корпустық лингвистикада негізгі назар тілдік жүйенің сөйлеу арқылы жүзеге асуына көп көңіл аударылады: «Как правило, в корпусной лингвистике языковой аспект фактически игнорируется, поскольку изначально фиксируется область привлекаемых данных – реализаций языковой системы» [10, 114-б.]. Солай бола тұра, корпусты құру мен оны түрлендіру жағын зерделеуде тілдік аспектіні ескеру қажеттігі күмән тудырмайды.

В.П.Захаровтың корпус ұғымын тар мағынада түсіндіруі қазіргі ғылыми түсінікте эволюциялық тұрғыда көрініс тапқан. Мысалы, ғалым «Корпусная лингвистика» атты оқу құралында былай дейді: «Под названием лингвистический, или языковой, корпус текстов понимается большой, представленный в электронном виде, унифицированный, структурированный, размечанный, филологический компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [18, 3-б.].

Корпуста сақталатын бірлік – ол пәндік саланың қайсыбір жиынтығы. Мысалы, оларға жататындар: сөз, сөзтіркес, сөйлем немесе бүтін мәтіндер жиынтығы [17]. Корпустағы деректер құрылымданған және құрылымданбаған пішімдерде сақталуы мүмкін. Біріншіде, яғни құрылымданған пішім жағдайында шартты белгілермен белгіленген мәтін туралы сөз болады. Ондай белгілер семантикалық (лексика-семантикалық), морфологиялық, синтаксистік сипатта болуы мүмкін. Мысалы, орыс тілінің ұлттық корпусында орын алған морфологиялық белгілер барлық сөздерді сөз табына, септелу категориясына, тегіне (род – муж., жен.), жіктелуіне, жанды-жансыздығына, етіс және етістік түріне, салыстырмалы шырай және т.б. морфологиялық сипаттамаларына қарай арнайы белгі қою жүргізілген. Сол сияқты семантикалық талдау арқылы да мәтін ішіне қажетті белгілердің қойылатынын айтуға болар еді. Мысалы, орыс тіліндегі үстеу сөздерге «Таксономияға», «бағалауға», «сөзжасамдыққа» қатысты белгілеулер орын алған. Мұндағы «Таксономия» белгісі: орынды, бағытты, қашықтықты, уақытты, жылдамдықты, санына және т.б. жүйелеулерге қатысты семантикалық сипаттамаларды білдіреді.

Корпустарға, сан жағынан, алғанда бірнеше талаптар қойылатыны да белгілі. Олардың негізгісі – репрезентативтілік (тұлғалылық), яғни оны сол корпустың пәндік аяның барлық қасиетін бейнелей алу мүмкіндігі деуге болады. Басқаша айтқанда, сол лингвистикалық зерттеу типіне қатысты пәндік аядағы құбылыстың кездесу жиілігі тілдік бірліктерді бір-бірінен ажырата алатындай мәнде болуы қажет [10, 118-б.]. Репрезентативтіктің бұл қасиеті сандық аспект ретінде қарастырылады да, ал оны зерттеп анықтау математикалық әдіспен жүзеге асады. Сонымен бірге репрезентативтіктің филологиялық әдіспен зерттелетін сапалық жағы да орын алады. Ал ондай сапалықты корпустағы мәтіндер жиынтығының жанрлар мен функционалдық стильдермен теңдестік сипатта болуы деп ұғыну қажет. Басқаша айтқанда, тақырыптас мәтіндер бойынша жүргізілген статистикалық зерттеу нәтижелерінің ақиқаттыққа жақындығына әрі таңдама мәтін саны, әрі олардағы авторлар саны негіз болады.

А.Н.Барановтың тұжырымдауынша репрезентативтікті жүзеге асырудың ең қарапайым жолы – пәндік салалар бойынша алынатын таңдама мәтіндерді сәйкес деңгейде шағындау әдісі. Міне, тек осы жағдайда ғана корпусты ұйымдастыруда сәйкестік стратегиясы сақталатын болады [10].

Корпусқа қойылатын басқа да талаптар бар, олар мыналар:

– релевантты құбылыстардың толық ескерілуі (тіпті пропорционалды шағындауға сәйкес келмесе де);

– корпустың үнемділік қасиеті пәндік аяны зерттеушінің де іс-әрекетін үнемдеуі;

– тілдік матералдарды құрылымдау жолының пайдаланушы үшін де аса маңыздылығы;

– корпус деректерін өңдеуге қажетті компьютерлік бағдарламамен қамтамасыз етілуі (статистикалық талдау, конкорданс құрастыру, сөздікті

автоматты өңдеу – толық және ішінара сөзтізбелер алу: әліпбилі, жиілікті, лемматизациялау әрекеттері) [10].

Корпустарды келесі топтарға бөліп қарастыруға болады:

I. Құрылымдық дәрежесіне қарай:

1. Электрондық архив (сыртқы есте сақтау таспадағы өңделмеген мәтіндер);

2. Электрондық кітапхана (өңделген пішіндегі мәтіндер);

3. Мәтіндер корпусы (мәтіндердің стандартталған, өңделген және әр пәндік саланы бейнелейтін жиынтығы).

II. Тілдердің қамтылу санына байланысты:

1. Біртілді. 2. Екітілді. 3. Көптілді.

Сондықтан параллель корпустар дегеніміз – екі және көптілді корпустардың түрлері.

III. Қолдану тәсіліне қарай:

1. Зерттеу корпустары (көбінесе тілдік жүйелер қызметінің әртүрлі аспектілерін зерттеуге арналған).

2. Иллюстративтік (көрнекілік) корпустары (олар зерттеу жүргізілгеннен кейін ғана құрылады; бұл корпустардың мақсаты – алынған нәтижелерді растау және негіздеу). Мұндай қарама-қарсы қоюшылық корпустардың ерте даму кезеңіне ғана тән болғаны белгілі, себебі қазіргі кездегі корпустар әрі зерттеу жүргізу үшін, әрі көрнекілік үшін де қолданылуы мүмкін.

IV. Динамикалық сипатына қарай:

1. Тілдік жүйенің белгілі қалпын бейнелейтін статикалық корпустар.

2. Динамикалық (мониторлық) корпустар. Олар пәндік аяның процессуалды аспектісіндегі тілдік феномендердің белгілі уақыт ішіндегі қызметін айқындау мәселелерін қарастырады.

V. Индекстеуге қатысты:

1. Белгіленбеген жай корпустар.

2. Белгіленген (аннотацияланған, мазмұнданған) корпустар. Қазіргі кездегі корпустар тек аннотацияланған болып келеді, себебі тек белгіленген корпустардан ғана қажетті ақпаратты іздестіруге, пайдалануға мүмкіндік бар.

VI. Тіл формасына (тұрпатына) қарай:

1. Жазба түрдегі корпустар.

2. Сөйлеу түріндегі (фонетикалық) корпустар.

Корпустық лингвистиканың жетістіктерін өзіне сақтаған аса дамыған корпус түрі – ұлттық корпус. Мұндай корпус белгілі дәрежеде ұлттық тілді толық түрде бейнелейді. Ұлттық корпустың репрезентативтілігі (тұлғалылығы) – сол тілдің жазба және сөйлеу түріндегі мәтіндерінің барлық типтерінің бейнеленуі. Ұлттық корпустың айтарлықтай дәрежеде көлемді (ондаған, жүздеген миллион сөзқолданыс) болуы репрезентативтілікке жетудің қажетті шарты болып саналады. Ұлттық корпустың ажыратылмас бөлігі оның белгіленген (аннотацияланған, мазмұндалған) бейнесі.

Ұлттық корпус ғылыми зерттеулердің түр-түрін жүргізуді қамтамасыз етеді: лексикографияға, жасанды интеллектіге, әдебиеттануға, сөйлеу тілін

талдау мен жинақтауға және лингвистиканың барлық салаларына қатысты зерттеу түрлері. Сонымен бірге беделді академиялық сөздіктер мен ғылыми грамматикаларды құрастыру да корпустар негізінде жүзеге асады. Ұлттық корпусты пайдаланушылар – әртүрлі саладағы тілшілер, әдебиеттанушылар, тарихшылар және гуманитарлық білім салаларының өкілдері. Ұлттық корпустың ана тілі мен шет тілін оқытуда, оқулықтар мен бағдарламалар құрастыруда маңыздылығы аса зор деуге болады [17; 18; 19].

Қатар тілдер (параллель) корпустары бір тілден екінші тілге аударма жасауға қатысты талдау жұмыстарын жүргізуге аса қолайлы болып келеді. Мысалы, «Орыс тілінің ұлттық корпусында» параллель мәтіндердің (қатар тілдер мәтіндері) корпустары да орын алған. Мұндай корпустар ерекше корпустар қатарына жатады. Себебі, орыс тіліндегі мәтінге оның басқа тілге аударылған үлгісі және, керісінше, шет тілдеріндегі мәтіндерге орысша аудармасы сәйкестендірілген.

Түпкі және аударма мәтіндердің бірліктері арасында «теңестіру» нәтижесінде арнайы қарастырылған шаралар бойынша сәйкестік жүзеге асады. Теңестірілген паралельді корпус – ол ғылыми зерттеулердің, әсіресе, аударма жасаудың теориясы мен практикасының аса тиімді құралы.

Бүгінгі таңда тілдік сипаттамаларды теориялық тұрғыдан ғана негіздеп қоймай, сол тілдік жүйедегі заңдылықтар мен қағидаттарды, тұрақты категориялық сипаттамаларды автоматтандыру ісімен байланысты қарастыру қажеттігі дау тудырмайды. Бұл әсіресе, бүгінгі дамығын мемелекетіміздің ғылым мен техниканы үйлестіру мақсатын көздейтін негізгі міндеттерінің бірі.

Мәтіндерге лингвистикалық (грамматикалық) ақпараттар беру, корпустық зерттеулер біраз уақыттан бері жалпы тіл білімінде, соның ішінде шетел тіл білімінде лингвистикалық зерттеулердің негізгі тәсілдерінің бірі ретінде қолға алынып келеді.

Жоғарыда аталған зерттеу жұмыстары **қазақ тіл білімінде** бұрын-соңды зерттеу нысанына іліне қойған жоқ, тіпті орыс тіл білімінің өзінде де бұл мәселе лингвистиканың басқа салалары мен бағыттарына қарағанда кейіндеп қалған болатын.

«Мәдени құндылықтар ретіндегі қазақ тіліндегі мәтіндер корпусы және сөздіктердің «Тіл – қазына» атты ұлттық компьютерлік қоры» атты тақырып бойынша 2009 жылдан бастап жүргізіліп жатқан ғылыми-зерттеу жұмысының негізгі мақсаты – мәдени құндылықтар ретіндегі әлемдік дәрежедегі компьютерлік мәтіндер корпусын жасаудың лингвистикалық негіздемесін жасау. Ол үшін әлемдік дәрежедегі компьютерлік мәтіндер корпусын жасаудың лингвистикалық деңгейінің теориялық мәселелеріне байланысты материалдарға шолу жасау қажет болды. Сондықтан да компьютерлік лингвистика ғылыми тобының қызметкерлердің алдында, ең алдымен, мынадай міндеттер тұрды:

- тілдік корпус терминінің мәнін айқындап, оның лингвистикалық табиғатына сипаттама беру;
- тілдік корпус жасаудың теориялық негіздемелерімен танысу;

- корпус жасаудың лингвистикалық деңгейін компьютерлік (инженерлік) деңгеймен байланыстыра отырып, олардың арақатынасын анықтау (көркем тіл мен техника тілін орайластыру);

- мәтіндік бірліктеріне белгі-кодтарды сәйкестендірудің кезеңдік әдіс-тәсілдерін меңгеру, яғни лингвистикалық белгі-кодтарды қоюдың теориялық және практикалық мәселелерін шешу;

- морфологиялық талдаудың дәстүрлі принциптеріне сүйене отырып, талдау үлгілеріне шолу жасау;

- олардың ең тиімді, соңғы ғылыми жаңалықтармен байытылған нұсқаларын таңдап алу (сөз табы, сөзформалардың құрамына қарай дара және күрделі, деректі, дерексіз, жұрнақ, жалғау т.б. морфологиялық (грамматикалық) белгілерді сұрыптау) жұмыстары атқару.

Сонымен, 2009 жылы таңдама мәтіндер бойынша тиісті сөзформаларға лингвистикалық ақпарат берудің (белгі-код қоюдың) бірінші кезеңі, яғни мәтінде қамтылған сөзформалардың барлығына жаппай морфологиялық белгі-код қою, ал 2010 жылғы екінші кезеңде жаппай қойылған белгі кодтардың омонимдік қатарларын табу мен омоним сөзформалардың лингвистикалық (морфологиялық) сипаттамасын жасау көзделді.

Осы аталған жұмыстармен бірге инженерлік лингвистика тұрғысынан орындалатын, яғни компьютерлік базаның қажетті алгоритмін және жасанды тілдегі бағдарламсын жасау әдіс-тәсілдері де зерттелді.

«Мәдени құндылықтар ретіндегі қазақ тіліндегі мәтіндер корпусы және сөздіктердің «Тіл – қазына» атты ұлттық компьютерлік қоры» атты ғылыми-зерттеу тақырыбының теориялық және практикалық маңызына қысқаша тоқтала кетуді жөн көріп отырмыз.

Қазақ тілінің мәтіндік корпусын жасау мәселесінің шешілуі теориялық тұрғыдан да, практикалық тұрғыдан да құнды болмақ және мұндай зерттеу жұмысы қазақ тіл білімінде бұрын-соңды қолға алынбағандықтан, біріншіден, зерттеу нәтижесі компьютерлік лингвистика бойынша алғашқы теориялық тәжірибелердің негізін қаласа, екіншіден, тілдік материалды модельдеу нәтижесінде көптеген тілдік ақпараттарды алып, оларды тіл білімінің кез келген саласында нақты тілдік дерек (факт) ретінде пайдалануға, ұстануға мүмкіндік жасайды. Әсіресе, грамматикалық теориялардың жаңа қырынан ашылуына септігін тигізеді деп білеміз.

Қазақ тілінің тілдік корпусын зерттеуге арналған бұл жұмыстың нәтижелері жоғарғы оқу орындарында оқытылатын «Қазақ тілінің функционалды грамматикасы», «Компьютерлік лингвистика», «Морфология», «Мәтін лингвистикасы» т.б. салалар бойынша дәрістер мен семинар сабақтарында барынша пайдалануға болады. Ал аталған зерттеу жұмыстарының нәтижелері ретінде мына жайттарды атауға болар еді:

- әлемдік дәрежедегі компьютерлік мәтіндер корпусын жасаудың лингвистикалық деңгейінің теориялық мәселелеріне байланысты материал жинастыру және осы мәселелер бойынша жазылған ғылыми еңбектерді зерттеу, мақалалар жариялау. Міне, осы аталған іс-әрекеттердің өздері зерттеу ісінің алғашқы жаңалықтары деуге болады;

- тілдік корпус терминінің мәнін айқындап, оның лингвистикалық табиғатына берілген сипаттама нәтижесі;

- тілдік корпус жасаудың айқындалған теориялық мәселері;

- корпус жасаудың лингвистикалық деңгейін компьютерлік (инженерлік) деңгеймен байланыстыра отырып, олардың анықталған арақатынасы және айқындалған белгі-код қоюдың кезеңдік әдіс-тәсілдері;

- лингвистикалық белгі-кодтарды қоюдың қазақ тілі үшін шешімін тапқан теориялық және практикалық мәселелері.

Сол сияқты, морфологиялық талдаудың дәстүрлі принциптеріне сүйене отырып, талдау үлгілеріне шолу жасап, олардың ең тиімді, соңғы ғылыми жаңалықтармен байытылған нұсқаларының таңдалып алынуы да (сөз табы, сөзформалардың құрамына қарай дара және күрделі, деректі, дерексіз, жұрнақ, жалғау т.б. морфологиялық белгілердің сұрыпталуы) қазақ тіліне қатысты бұрын-соңды қарастырылмаған мәселелер.

Морфологиялық белгі-кодтар қоюдың принциптері айқындалғаннан кейін, мәтіндер корпусын жасау мақсатында Абай Құнанбаевтың толық шығармалар мәтініндегі, М.Әуезовтің «Абай жолы» романы және Ә.Кекілбайұы шығармалары және т.б. мәтіндеріндегі сөзформаларға морфологиялық талдаулар жасалып, оларды таңбалаудың шартты белгілері біріздендірілді. Мәтіндік корпус грамматикалық сөздік мәселесімен тығыз байланыста қарастырылатындығын да айта кетуге болады. Қазақ тіліндегі үлкен көлемді мәтіндер бойынша компьютерлік корпусты әлемдік деңгейде құрастыру үшін таңдаудан өткен алгоритмдік құрылым мен зерттеу мақсатына сай құрастырылған (сызылған) компьютерлік бағдарламаның сызба-топтамалары іске қосылып, сыннан өтті.

Мәдени құндылықтар ретіндегі қазақ тіліндегі компьютерлік мәтіндер корпусын құрастырудың лингвистикалық деңгейі бойынша М.Әуезовтің 20 томдық шығармалар мәтініндегі және т.б. толық шығармалар жинақтарынан алынған таңдама мәтіндеріндегі сөзформаларды орфографиялық сөздерге бөліктеудің, сөздік бірлігіне келтірудің (лемматизациялаудың) және морфологиялық талдау арқылы оларға шартты белгі-кодтар қою әрекеті жүзеге асуда.

Қазақ тілін зерттеуде корпустық лингвистика саласына көп мән берілуі және оның әлемдік дәрежедегі теориялық және практикалық жақтарын зерттеу қажеттігі туындауда. Әлем бойынша мәтін корпустарын құрастыру мен олардың қызметіне қатысты жалпы және нақты мәслелелерге арналған мақалалар жарық көрген ғылымы журналдардың арнайы басылымдары да шыға бастады [12; 13]. Бірақ, әлде де, қазақ тіл білімі үшін корпустық лингвистикаға қатысты көптеген мәселелер арнайы зерттеуді қажет ететіні белгілі. Оған жататындар: корпустық лингвистика мен оның негізгі ұғымдарының анықтамалары, корпустық лингвистиканың тіл білімі құрылымында алатын орны, әдіс-тәсілдері және т.б. Сонымен бірге жаңа бағыттың теориялық негізін ұғыну мәселесі корпустарды нақты зерттеулерде пайдалануға қарағанда белгілі дәрежеде қалыс қалып келе жатқаны да байқалады.

Корпустық лингвистика пәнін осы саланың мамандары тілдік корпустарды құру мен оны пайдалану жағдайын зерттейтін тіл білімінің бір бөлімі ретінде ғана қарастырып келді. Кейбір ғалымдар ол пәннің түсінігін тар шеңберде қарастырып, оны тек компьютерлік лингвистика саласының аясында ғана түсіндіреді: «Корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с использованием компьютерных технологий» – дейді [14]. Ал компьютерлік лингвистика ұғымын, әдетте, компьютерлік құралдарды пайдаланудың кең мүмкіндігі ретінде түсіндіруге болатыны белгілі. Бұл жердегі «компьютерлік құралдар» деп отырғанымыз – компьютерлік бағдарламалар, тілдік деректерді өңдеу мен компьютерлік технологияны орынды ұйымдастыру жұмыстары және т.б. [15, 13-б.]. Ал корпустық лингвистика компьютерлерді тек «құрал» ретінде пайдаланатыны да бәрімізге аян. Міне, сондықтан да корпустық лингвистика өзіне жүктелген міндетті ондай құралсыз атқара да алмас еді. Бірақ компьютер мұндай рөлді қазіргі білім саласының барлық түрлерінде де атқаратынын ескерсек, онда олардың бәрін бірдей компьютерлік лингвистика саласына жатқыза беруге болмайды.

Жоғарыда сөз болған корпустық лингвистика пәнінің теориялық және тәжірибелік жақтары қазақ тілі мәтіндері бойынша компьютерлік корпустар базасын құру жағдайында да ескерілуі қажет. Корпустық лингвистика қазақ тіл білімінің де ерекше саласы ретінде қалыптасатын болса, қазақ тілшілеріне көлемді тәжірибелік материалдарды пайдалануға, қажетті деген тілдік деректерді тауып алуға және оларға тиісті деген өңдеулер жүргізуге мүмкіндік туындатады. Осының бәрі қазақ тіліне қатысты зерттеулердің шынайылыққа (ақиқаттыққа) жетудің эмпирикалық тәсілдеріне жаңаша көзқараспен қарауға және ғылыми айналым аясына аса маңызды тілдік материалдарды енгізуге жағдай жасайды.

Қазіргі кезде әлемдік корпустық лингвистиканың даму сипаты – ұлттық толық мәтіндерді арнайы зерттеу нысаны етіп алу. Сондықтан автоматтанған қазақ тіліндегі мәтіндер корпусының компьютерлік базасы (теориялық және практикалық жағынан қарастырғанда) жақын болашақта жүзеге асатын «Қазақ тілінің ұлттық корпусының» аса маңызды бастамасы болары сөзсіз. Мұндай зерттеулердің нәтижелері қазақ мәтіндерінің стилдік, құрылымдық, мағыналық, функционалдық және т.б. сипаттарын анықтауда да өзекті мәселелердің бірі болып саналады.

Қорыта айтқанда, компьютерлік лингвистика мамандарының тұжырымдауынша, компьютерлік тілдік қор дегеніміз ғылым адамының өз зерттеу нысанына жаңаша тұрғыда көз салу мүмкіндігі. Мұндай тілдік қор неғұрлым қомақты болса, солғұрлым тіл құрылысының сыры тереңірек ашылады, сөйтіп зерттелетін нысан жөніндегі түсініктердің аумағы кеңиді, адамның білім өрісіндегі «ақтаңдақтардың» бедер-бейнесі айқындала түседі. Сол сияқты, зерттеуші адамның қалып-қабілеті әлденеше есе артады, шығармашылық қуат көздері ашыла түседі, сөйтіп бұл жаңа мүмкіндіктер қазақ тілінің жүйелілік қасиеттерін жетілдіруге және тіл жүйесін мұқият

тануға жұмсалатыны сөзсіз.

ӘДЕБИЕТ

1. *Назарбаев Н.* Қазақстан жолы. –Қарағанды, 2006. –372 б.

2. Қазақ тексінің статистикасы. Статистика казахского текста. – Алматы: Қаз ССР-ның «Ғылым» баспасы, 1973.- 731 бет.

3. *Бектаев К.Б., Пиотровский Р.Г.* Математические методы в языкознании. Ч.1. Теория вероятностей и моделирование нормы языка, Алматы, 1973. 281 с.; Ч. 2. Математическая статистика и моделирование текста. Алматы, 1974. 334 с.

4. *Бектаев К.Б.* Статистико-информационная типология тюркского текста. – Алма-Ата, «Наука» КазССР, 1978. 167 с.

5. *Құрышжанов А.Қ., Жұбанов А.Қ., Белботаев А.Б.* Куманша-қазақша жиілік сөздік. – Алматы: ҚазССР «Ғылым» баспасы, 1978. 277 бет.

6. *Бектаев Қ.Б., Жұбанов А.Қ., Мырзабеков С. Белботаев А.Б.* М.Әуезовтің "Абай жолы" романының жиілік сөздігі . Алматы, 1979. 335 б.

7. *Бектаев Қ.Б., Жұбанов А.Қ., Мырзабеков С. Белботаев А.Б.* М.О.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері. Алматы-Түркістан: 1995. - 346 С.

8. *Жұбанов А.Қ.* Основные принципы формализации содержания казахского текста. – Алматы, 2002. – 250 с.

9. *Жұбанов А.Қ.* Қазақ сөзінің «**Тіл – қазына**» деректер базасы және оның теориялық негіздері. База данных «**Тіл – қазына**» казахского слова и ее теоретические основы. – Алматы: «Арыс» баспасы, 2009. – 304 бет.

10. *Баранов А.Н.* Корпусная лингвистика // Баранов А.Н. Введение в прикладную лингвистику: Учебное пособие. –М.: Едиториал УРСС, 2003. С. 112-137.

11. *Захаров В.П.* Корпусная лингвистика: Учебн.-метод. пособие. –СПб., 2005.48 с.

12. Научно-техническая информация. Серия 2. Информационные процессы и системы. 2003. №6.

13. Научно-техническая информация. Серия 2. Информационные процессы и системы. 2003. №10.

14. *Захаров В.П.* Корпусная лингвистика [Электронный документ] // http://download.yandex.ru/class/zakharov/CL

15. *Баранов А.Н.* Компьютерная лингвистика // Баранов А.Н. Введение в прикладную лингвистику: Учебное пособие. –М.: Едиториал УРСС, 2003. С. 13-38.

16. *Фрэнсис У.Н.* Проблемы формирования и машинного представления большого корпуса текстов // Новое в зарубежной лингвистике. Выпуск XIV. Проблемы и методы лексикографии. –М.: Прогресс, 1983. С.334-352.

17. *Вербицкая Л.А., Казанский Н.Н., Касевич В.Б.* Некоторые проблемы создания национального корпуса русского языка // Научно-техническая информация. Серия 2. 2003. №6. С.2-8.

18. *Шаров С.А.* Представительный корпус русского языка в контексте мирового опыта // Научно-техническая информация. Серия 2. Информационные процессы и системы. 2003. №6. С.9-18.

19. Национальный корпус русского языка // http://www.ruscorpora.ru.

I БӨЛІМ

ҚАЗАҚ МӘТІНІНІҢ КВАНТИТАТИВТІ ҚҰРЫЛЫМЫ

***

КВАНТИТАТИВНАЯ СТРУКТУРА КАЗАХСКОГО ТЕКСТА

ЧАСТОТНЫЕ СЛОВАРИ И ИНФОРМАЦИОННЫЕ ХАРАКТЕРИСТИКИ ЯЗЫКА

Мин-во выс. и сред. специального образ-ия Каз.ССР. Чимкентский Гос.пед.институт. Тезисы докладов научно-теоретической конференции профессорско-преподавательского состава института, посвященной 50-летию Великой Октябр. Соц. Рев-ции. Чимкент-1967. - С. 107-108. (Соавт.: К.Бектаев и др.).

Разрешение теоретических и прикладных задач лингвистики во многом зависит от постановки вопроса описания языка. Существуют различные подходы описания языка, и они зависят от поставленной цели. В этом аспекте определенное место занимает статистическое описание языка, опирающееся на достижения математических наук, в частности: теории вероятностей, математической статистики и теории информации.

Статистическая структура языка (модель) прежде всего опирается на подсчет лингвистических единиц, т. е, на составление частотных списков и распределений.

Составление частотных словарей, словосочетаний необходимо для прикладной лингвистики (автоматического перевода с одного языка на другой, оптимального кодирования и др.) и для собственно лингвистических исследований.

В результате многолетнего исследования авторами получены следующие данные:

1. Составлен полный частотный словарь языка великого казахского поэта и просветителя Абая, объемом в 46.847 словоупотреблений.

2. Составлен частотный словарь қазахской публицистики объемом в 150.000 словоупотреблений.

3. Составлены частотные словари русского, английского и немецкого подъязыков «физическая химия» объемом по 100.000 словоупотреблений.

4. Составлен частотный словарь трехсловных сочетаний русского подъязыка электроники, в которых в качестве ядра (опорной словоформы) в словосочетаниях были использованы:

1) существительные, 2) глаголы и категория состояния, 3) наиболее часто употребительные словоформы частотного списка объемом по 100.000 трехсловных сочетаний по каждому виду.

5. Впервые составлены на ЭВМ Минск-22 алфавитно-частотные и частотные словари по казахскому, узбекскому, туркменскому, азербайджанскому, киргизскому и каракалпакскому языкам, а также орхонской письменности объемом от 2 до 10.000 словоупотреблений по каждому языку.

Чимкентская подгруппа «Статистика речи», которая входит, как неотъемлемая часть общесоюзной группы, в «Статистику речи», руководимая доктором филологических наук, профессором Р.Г.Пиотровским, проводит еще исследования по определению информационных характеристик языка.

По методу угадывания, предложенному К. Шенноном, членами группы проведены экспериментальные работы и получены информационные характеристики:

1. Казахского языка на основе 60 экспериментов, проведенных по полной программе, 60 экспериментов по сокращенной программе, (стиль – проза и публицистика).

2. Узбекского языка на основе 50 экспериментов по полной программе (стиль – проза).

3. Греческого языка на основе 27 экспериментов по комбинации полной и сокращенной программы (стиль – публицистика).

4. На основе 30 экспериментов по каждому языку определены информационные эквиваленты казахского, русского и английского языков, за оригинал которого взят роман М. Ауэзова «Путь Абая» и его перевод на русский, с русского на английский язык.

К ВОПРОСУ ОБ АВТОМАТИЗАЦИИ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ

Материалы первой научной конференций молодых ученых АН КАЗ. ССР. Изд. «НАУКА» Каз. ССР. Алма-Ата, 1968. – С. 442-443.

1. Исследования по автоматической обработке и хранению словесной информации ставит лингвистику перед необходимостью обращения к современной электронной технике в качестве вспомогательного средства при

анализе языкового материала. Как известно, языковеды, занимающиеся лексикологией и лексикографией и другими разделами лингвистики, просматривают большое количество языкового материала, выписывают слова на карточках и затем сортируют их по заданным признакам. Подобное многократное повторение простых операций занимает значительную долю времени, необходимого для составления словаря. Столь трудоемкая работа в принципе может быть легко разрешена с помощью электронно-вычислительных машин (ЭВМ).

2. ЭВМ выполняет такие операции, как классификация слов по алфавиту, по числу букв, составу слогов, подсчет любых статистических параметров – среднюю длину слова или фразы, распределение грамматических категорий, сочетаемость слов, их грамматических информации и причем ЭВМ решает те лингвистические задачи, которые выразимы в виде последовательности логических операций, но громоздки по количеству обрабатываемого материала, как, например:

а) составление частотных словарей;

б) составление алфавитных словарей (обычных и обратных);

в) составление словоуказателей к текстам, т. е. конкордансов;

г) количественные исследования стиля данного автора, сравнение текстов;

д) оценка родства языков;

е) исследование вероятностного изменения языковой модели во времени и т. д.

3. В нашей стране центры по автоматизации лексикологических и лексикографических исследований начинают развиваться с недавних пор. Ныне по указанным выше проблемам работает большой коллектив статистики речи руководимый доктором филологических наук, профессором Раймондом Генриховичем Пиотровским (Ленинград), который объединяет около 70 лингвистов, математиков и программистов-инженеров.

Перед коллективом поставлена цель: исследовать различные аспекты языка точными методами, создать статолингвистические модели языка и их описания, накопить материал для нужд информационно-поисковых автоматических устройств, машинный перевод с одного языка на другой.

4. Упомянутая группа, членом которой являюсь и я, на ЭВМ «Минск-22» впервые осуществил экспериментальные частотные и алфавитно-частотные словари по казахскому, узбекскому, туркменскому, азербайджанскому, киргизскому и каракалпакскому языкам, а также орхонской письменности объемом от 2 до 10 тыс. словоупотреблений по каждому языку. Словари казахского текста объемом 5 тыс. и 10 тыс. словоупотреблений составлены по роману М. О. Ауэзова «Абай жолы» («Путь Абая»). Каждое словоупотребление имеет свою индексацию по частям речи и некоторые статистические характеристики, а именно: абсолютная частота, абсолютная накопленная частота, относительная и относительно-накопленные частоты, информационный вес каждого слова

$$H_i = -P_i \cdot \log_2 P_i,$$

где P_i – относительная частота словоформы, H_i накопленный информационный вес группы слов.

Для получения словаря была использована программа, составленная членами нашей группы А.В.Зубовым и Э.Н.Хотяшовым (с соответствующими для казахского языка изменениями).

5. Нами лично составлена программа для ЭВМ «Минск-22» и получены по тексту романа М.О.Ауэзова «Абай жолы» (объемом 10 тыс. словоупотреблений) списки всех предложений с высокочастотными словоформами, «триад» с высокочастотными глаголами и список предложений с определенным сочетанием частей речи.

Данные эксперименты носят как бы разведочный характер и отчасти связаны с необходимостью подготовки достаточно сложных аналитических программ для наших дальнейших работ.

АВТОМАТИЗАЦИЯ НЕКОТОРЫХ ЛИНГВИСТИЧЕСКИХ ПРОЦЕССОВ

«Вестник» АН КазССР. Исследования молодых ученых.
Алма-Ата, – 1968. –№ 9. – С. 31-36. (Соавт. А.В.Зубов)

В настоящее время электронно-вычислительные машины находят применение в различных областях науки и техники.

В лингвистике также накоплены некоторые данные об использовании электронной техники в исследовании фактов. Опыт показывает, что главное преимущество применения ЭЦВМ при этом состоит в экономии времени и в освобождении человека от формальных, нетворческих действий.

Как известно, исследователи в области лексикологии и лексикографии занимаются просмотром большого количества языкового материала, выпиской их сочетаний на карточки и последующей сортировкой их по заданным признакам: приведение отдельных словоформ к одной парадигме, расстановка по алфавиту, объединение всех повторений одного слова, сортировка по частоте употребления и т. д. Столь трудоемкая работа может быть легко разложена на элементарные операции и реализована с помощью ЭЦВМ.

Таким образом, ЭЦВМ решает те лингвистические задачи, которые можно представить в виде последовательности логических операций, и которые достаточно громоздки по количеству обрабатываемого материала.

Предварительная обработка текста. При решении любых лингвистических задач с помощью ЭЦВМ текст, подвергаемый исследованию, необходимо закодировать, т. е. перевести его на язык, понятный машине. Для этого текст переносится на перфокарты или перфоленты. Например, в машине «Минск-22» для этой цели используются

перфолента и телеграфный аппарат СТА-2М. При нажатии клавиша аппарата происходит перевод буквенных символов в международный телеграфный код М-2. Каждый буквенный символ на пятипозиционной ленте выступает в виде комбинаций пробивок (отверстий) [1]. Если учесть, что каждая пробивка читается как «единица», а отсутствие пробивки – как «нуль», то любой буквенный символ на ленте является определенным числом в двоичной системе счисления.

Например, слово «батыр» (герой) будет иметь вид в двоичной системе счисления изобразится следующим образом:

10011 – Б
11000 – А
00001 – Т
10101 – Ы
01010 – Р

В памяти машины каждая буква, знак и цифра представляется двузначным восьмеричным числом, как например, А = 70, Б = 63, (•) = 07.

Как правило, на первом этапе исследования с помощью ЭЦВМ необходимо как-то отметить (заиндексировать) интересующий нас объект исследования и только после этого перевести текст на перфоленту.

Так, чтобы исследовать некоторые лингвистические закономерности частей речи в романе М.Ауэзова «Абай жолы» («Путь Абая»), над каждым словом текста был поставлен индекс, определяющий принадлежность этого слова к той или иной части речи. Индекс ставился после каждого словоупотребления через дефис. Например, Абай-З, келді-Г и т. д.

Эти же индексы служат для различения омонимичных слов при построении частотного словаря данного романа.

Каждую из частей речи мы обозначали определенной буквой русского алфавита, а именно: существительное – З, прилагательное – С, числительное – А, глагол – Г, местоимение – Е, наречие – У, служебные слова – К, междометие – О, подражательные слова – Л, модальные слова – Д.

ЭВМ «Минск-22» располагает возможностью вводить и выводить лишь только тексты, написанные русским или латинским алфавитом. Некоторые казахские буквы не имеют русских эквивалентов. Поэтому специфические казахские буквы Ә, Ө, Қ, Ғ, Ұ, Ү, І, Ң, Һ мы обозначали соответственно символами Э, О', К', Г', У+, У, 1, Н', Х'.

Алгоритм вывода текста предложений, содержащих определенные опорные словоформы. На основе описываемой ниже программы, разработанной нами, можно составить список всех предложений или триад (трехсловное сочетание) казахского и других текстов, содержащих заранее заданные опорные слова.

На принципиальной *блок-схеме 1* данной программы рассмотрим назначение и работу каждого из блоков.

В ***блоке 1*** осуществляется подготовка программы для ввода первой порции текста (ввод информации производится порциями, содержащими определенное количество слов или предложений).

Блок-схема 1

Блок предварительных операции
Ввод очередной пороции текста
Вызов очередной ячейки, содержащей текст
4
= 0?
Выделение (очередной) словоформы
Организация предложения
Поиск в предложении заданных
8
Найдено?
Печать предложения с заданным опорным
Очистка робочих ячеек и счетчиков
Печать числа предложений в порции и числа выбранных предложений
Останов. – конец порции

Происходит очистка ячеек, в которых будут фиксироваться число предложений в порции и число слов в предложении, а также ячеек, содержащих предложения.В ***блоке 2*** осуществляется ввод первой (а затем второй, третьей и т. д.) порции текста с перфоленты. Ввод производится в режиме группового ввода с указанием одного лишь начального адреса массива. Каждому символу (букве, цифре, знаку препинания, пробелу и т. д.) соответствует двузначное восьмеричное число. Следовательно, весь текст в машине представлен непрерывной цепочкой восьмеричных чисел.

Каждая ячейка памяти машины вмещает по 6 символов. Признаком конца порции служит помета «–0» (минус нуль), которая засылается программно в конец каждой очередной порции текста. Проверка правильности ввода исходной информации осуществляется путем двойного ввода.

Так как каждому символу текста соответствует определенная комбинация нулей и единиц, то все вводимые символы можно сложить между собой в двоичной системе счисления. Получается некоторое число, которое называется контрольной суммой массива. Эта сумма всегда должна быть постоянной для данного массива. После первого ввода перфолента возвращается в исходное положение и снова производится ввод того же массива информации. При этом полученная контрольная сумма сравнивается с первоначальной. При совпадении контрольных сумм управление передается на ***блок 3***. При несовпадении их машина останавливается. В таком случае нажатием кнопки «пуск» на пульте управления повторяют этап двойного ввода до совпадения контрольных сумм.

В ***блоке 3*** из введенного в память ЭЦВМ массива осуществляется вызов содержимого первой (а затем второй, третьей и т. д.) ячейки в одну из рабочих ячеек памяти и сравнивается с признаком конца порции, т. е. «–0». При совпадении управление передается на ***блок 11***, где осуществляется контрольная печать числа предложений в порции и числа напечатанных предложений с опорной словоформой. Затем вводится следующая порция текста, и весь цикл обработки повторяется.

В случае несовпадения с признаком конца порции управление передается на ***блок 5***, где осуществляется выделение словоформы. Слова могут состоять из одной, двух и более букв. Они отделены друг от друга пробелами, которые также являются символами. Поэтому выделение очередной словоформы осуществляется путем выявления последующего пробела, для чего необходимо сопоставление каждого символа с пробелом. Выделенные символы пересылаются последовательно в четыре ячейки памяти.

В ***блоке 6*** контролируется результат сравнения выделенных символов с пробелом, тогда какой-то выделенный очередной символ окажется пробелом. Это означает, что в упомянутых выше четырех ячейках памяти выделена очередная словоформа текста. Так как каждая ячейка памяти машины «Минск-22» может содержать не более 6 букв, то наибольшее допустимое число букв в словах обрабатываемых текстов не должно превышать 24 букв.

Алгоритм автоматического выделения предложений, содержащих определенные сочетания слов.

Блок-схема 2

В ***блоке 6*** содержимое четырех рабочих ячеек, содержащих выделенное слово, пересылается в заранее подготовленные ячейки памяти до тех пор, пока содержимым этих ячеек не окажется точка. Пересылкой точки в рабочее поле для предложения заканчивается процесс организации предложения. После того, как закончен процесс организации предложения, каждое слово его сравнивается поочередно с заданными опорными словами, которые заранее занесены в память машины. Если хотя бы одно из слов предложения найдется среди опорных, все предложение выпечатывается на алфавитно-цифровом печатающем устройстве (АЦПУ). В противном случае предложение пропускается, и машина переходит к чтению первого слова следующего предложения и т. д. (см. ***блоки 7, 8, 9, 10)***.

Приведенный выше алгоритм не использует всей той индексации, которая применена при кодировании текста. В то же время полная индексация текста открывает широкие возможности для автоматического синтаксического анализа [2], для статистического изучения сочетаний слов в предложении и их управление.

Рассмотрим простейшую задачу: как часто, и в каких предложених встречаются рядом сочетания прилагательного и глагола.

Алгоритм решения этой задачи во многом сходен с предыдущей задачей, в частности совершенно аналогичны блоки 1-6.

ФРАГМЕНТ 1

1. У’Ш-А КУ’Н’Д1К-С ЖОЛДЫН’-З БУ’ГІНГІ-С СОН’Г’Ы-С КУ'НІНЕ-З ШЭКІРТ-З БАЛА-З БАРЫН-З САЛДЫ-Г.
2. КУ'Н У+ЗЫН-У АТТАН-З ДА-К ТУ'СПЕЙ-Г ӨЗГЕ-Е ЖУ'РГІНШІЛЕРДЕН-З, ОК'-З БОЙЫ-К АЛДА-У ОТЫРГ'АН-Г.
3. ШЭКІРТ-З БАЛА-З У+ЗАП-Г КЕТКЕНДЕ-Г БУ+ЛАР-Е ДА-К ЕРІКСІЗ-С ЖЕЛЕ-Г ШОК'ЫРАК'ТАП-Г ШАУЫП-Г ОТЫРЫП-Г К'УЫП-Г ЖЕТЕДІ-Г.

ФРАГМЕНТ 2

1. ЖОК'-Д ЕКЕН-Г.
2. К'АУЫПТЫ-С ДЕГЕН-Г ЕСЕНБАЙГ'А-З
3. ДЕП-Г СОК'ТЫРАР-Г

ФРАГМЕНТ 3

1. ОСЫ-Е К'АЗІРГІДЕЙ-У БОП-Г ЖУ+РТ-З ШЫН'Г'ЫСК'А-З, ЖАЙЛАУГ'А-З К'АРАЙ-К КО'ШКЕНДЕ-Г ЕЛСІЗ-С БОП-Г К'АЛАТЫН-Г ЖЕР-З
2. К'ОЛДЫ-С БОЛСА-Г БОЛАТЫН-Г ЖОРГ’А-С ЖУ+МАБАЙ-З.
3. СОЛ-Е ДАЛАДАН-З КО'З-З АЛМАЙ-Г, ТОЯ-Г АЛМАЙ-Г, У’НСІЗ-С ТЕЛМІРІП-Г, У+ЗАК'-У + ЗАК'-У К'АРАЙДЫ-Г

В ***блоках 7–8*** из первого слова предложения выделяются два последних символа и сравниваются с признаком «-С» (т. е. прилагательным). Если да, то в блоках ***9–10*** выделяются два последних символа второго слова

предложения и проверяется, является ли это слово глаголом (-Г). Если – глагол, то предложение целиком выпечатывается.

При отрицательных результатах проверки в ***блоках 8*** и ***10*** управление передается на **блок *12***, где программа подготавливается для выделения символов (-С) из второго слова и (-Г) из третьего слова (затем из третьего и четвертого, четвертого и пятого и т. д.).

В ***блоке 13*** выделенные слова проверяются на совпадение с точкой. В случае совпадения это означает, что все предложение проверено и в нем нет указанного сочетания. В противном случае управление передается ***блоку 7***.

Как и в предыдущем алгоритме, после окончания обработки очередной порции текста машина останавливается. После установки новой порции процесс обработки повторяется.

Все эти циклы, повторяясь на каждой порции текста, осуществляют выделение из заданного массива этих и им подобных сочетаний.

Использование рассмотренных алгоритмов позволило получить на ЭЦВМ «Минск-22» по тексту романа М.О.Ауэзова «Абай жолы» списки всех предложений с высокочастотными словоформами «триад», с высокочастотными глаголами и список предложений с определенным сочетанием частей речи (фрагменты 1, 2, 3).

Таким образом, вышеописанные алгоритмы и конкретные данные, полученные на ЭЦВМ, представляют первую попытку применения автоматизации в тюркологии и открывают новые пути в тюркологических исследованиях.

Литература

1. *В. В. Шураков, В. В. Каплинский.* Программирование учетно-статистических задач для ЭЦВМ Минск-22. М., 1967.

2. Изучение синтаксических структур английского языка с помощью электронной машины. В сб.: «Автоматизация в лингвистике». М., стр. 144–150.

АВТОМАТИЧЕСКОЕ ПОСТРОЕНИЕ ЧАСТОТНЫХ СЛОВАРЕЙ (ПРЯМОГО И ОБРАТНОГО)

«Вестник» АН КазССР. 1970, № 3. – С. 48-53.
(Соавт.: К.Б.Бектаев, А.В.Зубов)

Применение в лингвистике новых, более точных методов анализа языковых явлений на базе структурного моделирования, статистического описания с использованием возможностей ЭВМ продиктовано временем. Оно опирается на выявление частотности языковых единиц (фонем, морфем, слов, словосочетаний и т. д.).

Языковеды Казахстана провели в этом направлении ряд экспериментальных исследований и получили весьма положительные результаты. Так, в 1968 году Институт языкознания АН Казахской ССР выпустил полный словник великого казахского поэта Абая Кунанбаева [1].

Словарь языка Абая является после словаря языка А. С. Пушкина единственным, в котором помимо толкования слов указывается адрес каждой словоформы, частота ее встречаемости.

Аналогичные словари (частотные, алфавитно-частотные) по казахской публицистике, по четырехтомному роману М. Ауэзова «Абай жолы» и другие составляются группой «Статистико-лингвистических исследований» Института языкознания АН КазССР, которая по своей проблематике входит в общесоюзную группу «Статистика речи», руководимую ленинградским профессором Р. Г. Пиотровским. Группа объединяет ленинградских, минских, казахстанских, молдавских, дагестанских и других ученых. В настоящее время в ней более 100 научных работников: языковедов, математиков, инженеров-программистов, работающих в трех направлениях: а) выяснение кодовых (энтропийных) свойств языка; б) определение статистических характеристик словаря, грамматики, фразеологии; в) автоматический анализ текста.

В Казахстане эти исследования ведутся на материале художественных, публицистических и научно-технических текстов [2]. На современном этапе развития лингвистической науки ощущается большая потребность в достоверном статистическом описании языка на различных его уровнях. В связи с выявлением частотности функционирования тех или иных языковых единиц в данной статье рассматриваются основные положения автоматического составления прямых и обратных частотных словарей с помощью ЭВМ.

Частотные словари и их роль в лингвистике. Частотным словарем называется упорядоченный список всех слов какого-либо текста. Под словом понимается любая цепочка букв между двумя пробелами. Список упорядочивается по частоте употребления слов. Первым ставится слово с наибольшей частотой, остальные слова располагаются за ним в порядке ее уменьшения. Слова, имеющие одинаковую частоту, упорядочиваются еще и по алфавиту.

Помимо этого составляется также обратный частотный словарь, где слова рассортированы по алфавиту, начиная с конца слова. При каждом слове указывается частота его употребления в том или ином конкретном тексте.

Материалы частотных словарей исключительно ценны как для собственно лингвистических исследований, так и для прикладных целей. Ими пользуются в решении основной проблемы общей лексикологии – выделение словарного фонда активного и периферического словаря; полезны они и при изучении вопросов стилистики и семантики [3].

Минимальное количество различных слов в речи повторяется с очень большой частотностью. Например, 135 наиболее часто встречающихся у Абая слов употреблено в 50% текста. Из этого можно составить представление о характерных особенностях использования лексических

средств поэтом. Слова, употребленные Абаем только по 1 и 2 раза, составляют 64% от всего словарного объема автора. Однако это количество слов по отношению ко всем словоупотреблениям составляет только 8%, что свидетельствует не только о лексическом богатстве казахского языка, но и о богатстве языка самого автора [4].

Особо важную роль могут сыграть частотные словари в построении методики преподавания и изучения иностранных языков. Они позволяют выделить наиболее употребительные единицы языка и на их основе составить учебники и учебные пособия по тому или иному языку.

В использовании счетно-аналитических и электронно-вычислительных машин в прикладной лингвистике накоплен определенный опыт как в нашей стране, так и за рубежом. Достижения комбинаторной (структурной) лингвистики стали основой при разработке программ автоматического перевода и реферирования печатного текста. Однако память современных кибернетических машин ограничена и не может вместить всех правил, которые определяют тот или иной естественный язык. Единицы языка, представляющего собой чрезвычайно избыточную систему, употребляются в частотном отношении неодинаково. Избыточность исследованных европейских языков доходит до 80%. Поэтому для машинного перевода особо важно составить частотные словари наиболее часто употребляемых и информационно эффективных слов [5].

В обратном словаре слова с одинаковыми слово- и формообразующими аффиксами располагаются рядом. Такой указатель чрезвычайно полезен для решения многих частных лингвистических задач. Так, он может использоваться как словарь рифм, т. е. служить базой для расшифровки древних текстов, давать ценный материал для анализа частотности аффиксов, особенно тюркских языков [6].

Ручное составление частотных словарей хотя и не сложно, но отнимает много времени и утомляет исследователя, что ведет к большому числу ошибок. Поэтому-то и возник вопрос об автоматическом построении частотных словарей. Первые опыты в этом направлении принесли положительные результаты [7].

Сотрудники группы «Статистика речи» при Минском государственном педагогическом институте иностранных языков, Институте языкознания АН КазССР, а также при Чимкентском педагогическом институте пользуются в своей работе электронными машинами. Основная часть работ выполняется на материале английского, французского, немецкого, испанского и других языков. Получены пробные материалы прямых и обратных частотных словарей различных языков тюркской группы (казахского, узбекского, киргизского, азербайджанского, туркменского, древнетюркского), составленных автоматическим способом [8]. Некоторые статистические данные этих словарей по тюркским языкам, полученные на ЭВМ «Минск-22», приведены в ***таблице 1***. Энтролия словоформ (H_i) вычислена по известной формуле $H_i = -\sum_i f_i = \log_2 f_i$, где f_i – относительная частота (вероятность) словоформ.

Таблица 1

Языки	Название источника	Объема текста	Количество разных словоформ	Средняя информ-я мера на единицу словоформ
Азербайджанский	Газетный текст	3000	1977	9,8967
Казахский	М. Ауэзов. Роман «Абай жолы»	5000	2532	10,4034
Казахский	М. Ауэзов. Роман «Абай жолы»	10000	4294	10,9159
Киргизский	Газетный текст	4000	2139	10,3834
Туркменский	Газетный текст	3000	1674	10,0647
Узбекский	Айбек. Роман «Навои»	3000	1917	10,3810
Орхонский	Памятник «Кюльтегин»	2000	882	9,0095

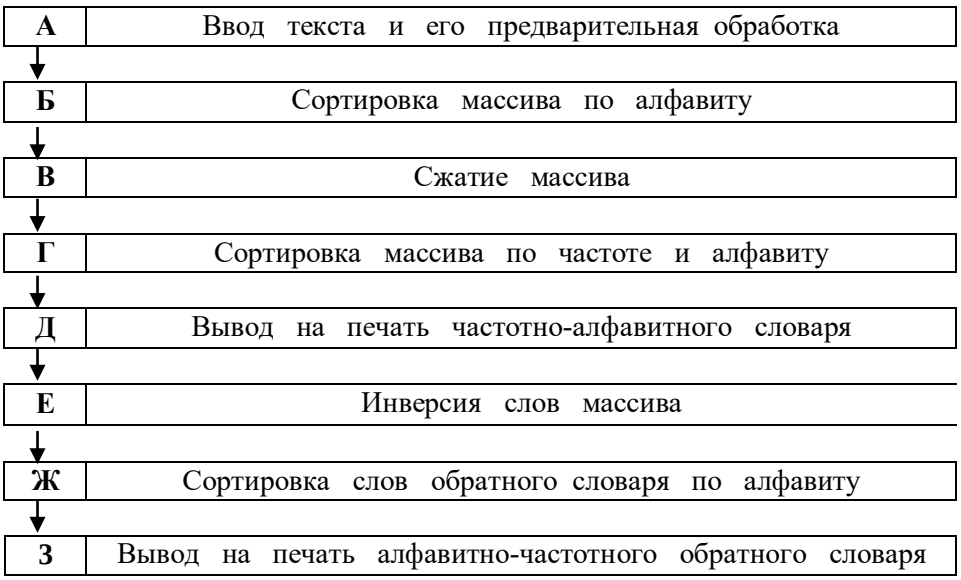
Чтобы получить из текста необходимую информацию с помощью ЭВМ, слова или группы слов снабжаются некоторыми индексами, по которым машина ведет анализ этих слов и их окружений. Так, заиндексированы в текстах все части речи (классы слов) и некоторые интересующие нас типы слов.

В текстах же тюркских языков встречаются буквы, которые отсутствуют в русском и латинском алфавитах. Поэтому нами разработана специальная система букв-заместителей.

Текст, подготовленный к автоматической обработке, в зависимости от типа используемой ЭВМ можно перенести на перфокарты или же на перфоленты [9].

Излагаемая ниже программа автоматического составления прямого и обратного частотных (и алфавитно-частотных) словарей по языкам тюркской группы может быть реализована для текстов сколь угодно большого объема.

Блок-схема алгоритма построения частотных словарей



Алгоритмы нашей задачи состоят из нескольких отдельных подалгоритмов, связь между которыми можно представить в виде укрупненной блок-схемы алгоритма построения частотных словарей, которая приводится ниже.

а) *Ввод текста и его предварительная обработка.* На этом этапе работы из каждой порции в 1000 слов отбираются отдельные слова. Подсчитывается частота их употребления в соответствующей порции. Каждая обработанная таким образом порция записывается во внешней памяти машины – на магнитных лентах (МЛ).

б) *Сортировка массива по алфавиту.* После того как весь массив, состоящий из *n* порций, записывается на МЛ, его необходимо рассортировать по алфавиту. Для этой цели нами была использована «Универсальная программа сортировки для ЭВМ «Минск-22», составленная Э.П.Хотящевым [10]. Так как задачи упорядочения лингвостатистических единиц во многом идентичны некоторым экономическим задачам, то мы имеем дополнительную возможность использования для целей лингвистического анализа готовых программ по решению экономических задач.

в) *Сжатие массива.* Так как в различных порциях могут встретиться одинаковые слова, то при сортировке всего массива по алфавиту они будут располагаться одно за другим вместе с их частотой. Например, если в первой порции слово «адам» (человек) встретилось 3 раза, слово «көз» (глаза) – 1, а слово «үлкен» (большой) – 1 раз, во второй: адам – 4 раза, көз–1 раз, үлкен–3 раза; в третьей: адам – 2 раза, көз – 3 раза, үлкен – 3 раза и т.д., то после сортировки по алфавиту эти слова расположатся так:

$$\left\{\begin{array}{lll} \text{адам} - 3 & \text{көз} - 1 & \text{үлкен} - 1 \\ \text{адам} - 4 & \text{көз} - 1 & \text{үлкен} - 3 \\ \text{адам} - 2 & \text{көз} - 3 & \text{үлкен} - 3 \end{array}\right.$$

Задачей алгоритма сжатия является сложение частоты одинаковых рядом стоящих слов и вместо многократной их записи запись каждого слова по одному разу в следующем виде:

адам – 9
көз – 5
үлкен – 7

г) *Сортировка массива по частоте и алфавиту.* После сжатия массива на МЛ окажется алфавитный список слов с указанием частоты встречаемости каждого слова в массиве из порций. Этот список необходимо рассортировать по частоте, т. е. сделать так, чтобы на первое место стало слово с наибольшей частотой употребления, за ним слово с меньшей частотой и т. д. Если ряд слов имеет одинаковую частоту, то их необходимо рассортировать еще и по алфавиту. Эта задача выполняется программой, описанной в пункте б).

д) *Вывод на печать частотно-алфавитного словаря.* Для вывода нами использовалась АЦПУ-128 (алфавитно-цифровое печатающее устройство) со скоростью 400 строк в минуту. На печать выдавались: номер слова, само слово, частота его встречаемости в выбранном тексте и абсолютная накопленная частота данного слова, являющаяся суммой частот всех предыдущих слов, и др. (***табл. 2***).

Таблица 2

Часть частотного словаря с информационной характеристикой (казахский текст)

№	Словоформа	Абсолютная частота	Накопл. абсолютная частота	Относительная частота	Накопл. относительная частота	H_I (энтропия)	H_I* (накопленная энтропия)
1	ДА–К	188	188	0,0188	0,0188	0,1078	0,1078
2	ДЕП–Г	144	332	0,0144	0,0332	0,0881	0,1959
3	ДЕ-К	121	453	0,0121	0,0453	0,0771	0,2729
4	БІР–А	117	570	0,0117	0,0570	0,0751	0,3480
5	ЕДІ-Г	96	666	0,0096	0,0668	0,0643	0,4124
270	ІНІСІ–З	5	4297	0,0005	0,4297	0,0055	3,7381
271	АГ'А–З	5	4302	0,0005	0,4302	0,0055	3,7435
430	СОНЫМЕН–Е	4	5001	0,0004	0,5001	0,0045	4,5223
431	СУ+Р–С	4	5005	0,0004	0,5005	0,0045	4,5268
4294	Я–О	1	10000	0,0001	1,0000	0,0013	10,9159

е) *Инверсия слов массива.* Из списка, полученного по алгоритму, описанному в пункте в), можно получить обратный частотный словарь путем операции инверсия. При этом слово оказывается записанным на МЛ в обратном порядке. Так, слово, имевшее до инверсии вид *үлкен,* после инверсии запишется как ***неклү***.

ж) *Сортировка слов обратного словаря по алфавиту.* После применения операции инверсия к каждому слову массива, записанному на МЛ, весь массив будет представлять собой обратный словарь тех же слов с указанием их частоты. Аналогично изложенному в пункте б) этот массив подвергается сортировке по алфавиту.

з) *Вывод на печать алфавитно-частотного обратного словаря.* Вывод на печать осуществляется точно по той же программе, алгоритм которой приведен в пункте д). Разница заключается лишь в том, что каждое слово перед печатью подвергается обратной инверсии с тем, чтобы на бумаге оно читалось слева направо.

Итак, по комплексной программе, разработанной группой «Статистика речи», на ЭВМ получают полные данные, начиная от составления частотных списков лингвистических единиц до вычислений, оценки статистических параметров, статистических проверок гипотез о распределении и т. д. Осуществляются опыты по переводу именных и глагольных сочетаний с немецкого, английского и французского языков на русский и другие языки народов СССР. Большую пользу приносит автоматизация лингвистических работ и для казахского прикладного языкознания в разрешении насущных задач лексикографии.

Литература

1. Абай тілінің сөздігі. Алматы, «Ғылым» баспасы, 1968
2. М.Н.Баркун, К.Б.Бектаев. Статистика речи. Перспективы и задачи

(Тезисы докладов). Алма-Ата, 1966.

3. Л.Н. Засорина. Частотные словари и вопросы лексико-статистики. Изд-во ЛГУ. 1966.

4. К. Б. Бектаев. Математика және лингвистика. «Білім және еңбек», № 6. Алматы, 1967. Его же. «Частотный словарь языка Абая». Изд-во ЛГУ, 1966.

5. Р. Г. Пиотровский. Информационные измерения печатного текста. Сб.: «Энтропия языка и статистика речи». Минск, 1966.

6. См.: (Автоматизация н лингвистика». Л., 1965.

7. П. М. Алексеев. Часготные словари как результат лексикостатистического анализа речи. ЛГУ, 1966.

8. К. Ф. Лукьяненков, Э. Н. Хотящев. Составление частотно-алфавитного списка с помощью ЭВМ. Сб.: «Вопросы романо-германской филологии и статистики речи». Минск, 1965; А. В. Зубов, Э.Н. Хотящев. Статистический анализ текста с помощью электронно-вычислительной машины. Сб.: «Энтропия языка и статистика речи». Минск, 1966.

9. Подробно см.: А.К.Джубанов, А.В.Зубов. Автоматизация некоторых лингвистических процессов. «Вестник АН КазССР», 1968, № 9.

10. Э.П.Хотящев. Универсальная программа сортировки для ЭВМ «Минск-22». Минск, 1965.

УНИВЕРСАЛЬНАЯ ПРОГРАММА ПОЛУЧЕНИЯ НА ЭВМ «МИНСК-22» АЛФАВИТНО-ЧАСТОТНОГО СПИСКА СОЧЕТАЕМОСТИ ГРАФЕМ

Материалы второй научной конференции
молодых ученых АН Каз.ССР. Алма-Ата,
1970. – С. 512-513.

Противопоставление объектов по дифференциальным признакам и исследование их диструбуции в контекстах является одним из основных приемов структурного анализа языка.

Нами проводятся работы по классификации множества графем казахского языка. Составили программу для ЭВМ «Минск-22», выявляющую частоты всречаемости и сочетаемости графем в тексте и словоформах, частоты длин предложений и словоформ. Программа рассчитана для ввода информации на русском и латинском шрифтах. Эта программа реализуется ныне в процессе статолингвистических исследований художественного текста длиной в 50.000 словоупотреблений, взятого из романа классика казахской литературы М.Ауэзова «Абай жолы».

Получены частотные списки графем по сплошному тексту, в начальных, средних и конечных позициях словоформ и алфавитно-частотные списки 2-х и 3-х графемных сочетаний в пределах текста и

словоформ. Сочетания графем обладают значительной степенью информацией о слове, о его частях и стыковых местах этих частей.

Различия между классами графем и их сочетаний позволят судить о слове по его наиболее информативным элементам и облегчает сам процесс описания принципов организации структуры письменного текста в соответствии с техническими возможностями воспринимающих устройств.

ИНФОРМАЦИОННОЕ И СТАТИСТИЧЕСКОЕ ИЗУЧЕНИЕ ТЮРКСКИХ ЯЗЫКОВ

Известия академия наук Каз.ССР,
№3, май – июнь. Изд-тво «Наука»
Каз. ССР, Алма-Ата, 1971. С. 82-84.
(соавторы: А.В.Зубов, С.Мирзабеков)

Семинар по информационному и статистическому изучению тюркских языков, организованный по инициативе Президиума АН КазССР, Института языкознания Академии наук республики и Всесоюзной группы «Статистика речи», состоялся 19–21 мая 1970 г. в Алма-Ате.

Вопросы, обсуждавшиеся на семинаре, поставлены самой жизнью. Нормализация языка и освоение литературного наследства классиков, а также исследование исторических памятников требуют просмотра миллионов страниц текста. Для этого нужна работа сотен языковедов. Например, 17-томный академический словарь русского языка составлен огромным коллективом более чем за 20 лет. Применив электронно-вычислительные машины, эти сроки можно было бы сократить в несколько десятков раз.

В тюркологии есть целый ряд задач, которые не могут быть решены традиционными методами. Но они с успехом решаются с помощью электронно-вычислительных машин (ЭВМ) на основе статистики и теории информации. Первые результаты, полученные в Институте языкознания АН КазССР, позволяют говорить о целесообразности применения таких методов.

Работа семинара проходила по секциям: «Частные вопросы статистических и структурных исследований тюркских языков» и «Автоматическая переработка текстов тюркских и других языков с помощью электронно-вычислительных машин и информационные измерения тюркских языков».

На пленарном заседании с докладом «О статистическом изучении тюркских языков» выступил К. Б. Бектаев (Институт языкознания АН КазССР). Отмечая самостоятельную теоретическую ценность статистических методов, он подчеркнул их большое значение в таких прикладных областях тюркологии, как дешифровка неясных мест в древнетюркских памятниках (орхонские и енисейские надписи), атрибуция анонимных тюркских текстов,

наконец, в исследованиях по культуре речи и по истории сдвигов и переоформления тюркской лексики.

Подчеркивая обширность материала, необходимого для получения достоверных данных, докладчик поделился первыми результатами, полученными с помощью ЭВМ «Минск-22» при составлении частотно-алфавитного, алфавитно-частотного и обратного словарей по языку романа М. Ауэзова «Путь Абая». Эту работу выполнил один человек за 8 месяцев.

В докладе Г. П. Мельникова (Москва) «О проблемах языка-эталона в тюркологии», посвященном нахождению общей точки («точки отсчета») группы языков или нескольких диалектов одного и того же языка, предлагается модель процесса появления группы родственных языков из одного языка-основы по аналогии распада (взрыва) движущегося по определенной траектории снаряда. Элементы, являющиеся результатом распада этого единого целого, и служат аналогами родственных языков или диалектов.

Оживленную дискуссию вызвал доклад доцента МГУ М. М. Глушко на тему: «Статистические методы в изучении структуры языка и текста». По мнению докладчика, частотные словари, составленные по словоформам, имеют ряд недостатков. Следовательно, для получения большой количественной информации необходимо использовать частотный словарь лексем. Он привел детально разработанную модель подготовки текстов для получения такого словаря, которая одновременно может дать и инвентари морфологических признаков лексем.

Доклады и сообщения, прочитанные на секции «Частные вопросы статистических и структурных исследований тюркских языков», относятся, во-первых, к статистико-информационным исследованиям внутри языков тюркской группы, во-вторых, к статистико-сопоставительному анализу языков тюркской и индоевропейской семей.

Ж. А. Аралбаев, К. Б. Бектаев, А. А. Исенгельдина, Б. Д. Тайлакбаев, С. С. Татубаев (Алма-Ата) в докладе «О статистическом исследовании фонетической структуры казахского языка» рассказали, как они, основываясь на системе перехода от графемного уровня к фонетическому, исследуют различные жанры современного казахского литературного языка. Применение электронно-вычислительных машин при этом дает возможность получить ряд таких важнейших характеристик казахской фонетики, как: а) частотность фонем и их вариантов по всему материалу и жанрам, б) частота дистрибуции фонем в слове и в слогах, в) структура различных типов слогов и их частотность.

О подробном исследовании, проведенном на материале узбекского языка, сообщали Е. Файзылов и А. Ризаев (Ташкент).

С. С. Татубаев (Алма-Ата) в докладе «Статистические подходы к исследованию певческой фонетики казахского языка» анализирует арии и ариозы из современных казахских опер, а также песни народных композиторов. Автор приходит к выводу, что благозвучие казахской речи снижается в результате выпадения конечной гласной, когда последующее

слово начинается с гласной, и за счет аллофонов *Қ* и *Ғ*, придающих гортанность речи.

А. Ахабаев, А. Белболатов, К. Молдабеков и Н. Осижигитова (Чимкент), исследовав казахский текст на уровне членов предложений, доказали, что структуры предложений научно-технических и художественных текстов казахского языка резко различаются.

На секции «Частные вопросы статистических и структурных исследований тюркских языков» заслушано 6 докладов.

М. М. Копыленко (Алма-Ата) в докладе «Описание структур казахского и русского языков в терминах аппликативной модели» пользуется аппликативной моделью С. К. Шаумяна. Приводятся примеры описания в терминах аппликативной модели своеобразия лексико-грамматического построения казахского и русского текстов.

Авторы доклада «Частотно-сопоставительные характеристики падежных форм русского и казахского языков» К. И. Ищанов, П. В. Садчикова, С. Мирзабеков (Чимкент), изучая тексты художественной литературы и философских произведений, установили значительное расхождение частот русского и казахского именительного падежа в аналогичных употреблениях.

Статистические характеристики форм страдательного залога в казахском и английском языках были даны в докладе А. Ергалауова (Алма-Ата). Так, разноструктурные английский и казахский языки имеют много общего в частотности употребления форм прошедщего, настоящего и будущего времени.

Р. М. Мустафина (Казань) в докладе «Статистико-информационное исследование перевода видов русского глагола на татарский язык» привела интересные данные, связанные с конфронтацией глагольных систем языков различной типологии.

Помимо перечисленных докладов, можно отметить сообщения Г. В. Ермоленко (Алма-Ата) «Статистика тюркизмов в романе И. Вазова «Под игом» (на материале болгарского языка») и А. Г. Байер, Н. С. Пак, Г. Г. Шауберг (Чимкент) «Статистика и преподавание иностранных языков» в тюркоязычной аудитории.

С большим интересом на секции «Автоматическая переработка текстов тюркских и других языков с помощью электронно-вычислительных машин и информационные измерения тюркских языков» был прослушан доклад профессора Р. Г. Пиотровского (Ленинград) «Многозначность и идиоматичность в машинном словаре». На основе анализа результатов многолетней работы коллектива «Статистика речи» по составлению машинных словарей докладчик определяет простую неоднозначность и идиоматичность через их статистические характеристики. Например, идиоматичные значения русского глагола «бить» и существительного «карман» (узб. «зиён етказмек», казах, «зыян істеу») можно рассматривать как редкие частные значения этих слов, обусловленные в первом случае правой дистрибуцией «по карману...», а во втором случае – левой

дистрибуцией «бить по ...». Анализируя проекты автоматического устранения многозначности и идиоматичности на материале европейских языков, автор высказывает предположение, что тюркско-индоевропейские машинные словари, если они когда-либо будут составлены, дадут значительно меньшую лексико-грамматическую неоднозначность и идиоматичность в связи с прозрачной и менее анонимичной морфологической структурой тюркских входных словоформ.

В совместном докладе А. А. Землянского, В. В. Морозенко, И. Ф. Турук, В. В. Шуракова (Москва) «Алгоритмизация лексико-статистического исследования письменной речи» рассматриваются вопросы выделения из текста лингвистической информации на уровнях частей речи и функциональных признаков. Практическим результатом исследователей является учебный словарь-минимум общенаучного слоя лексики английской научно-технической прозы и разработки критериев отбора словоформ. Предлагаемый ими алгоритм может быть применен и для тюркоязычных текстов.

Ю. Н. Марчук (Москва) в докладе «О дистрибутивно-статистическом определении классов слов» рассматривает две задачи: 1) разрешение последовательностей омографов в тексте и 2) дистрибутивно-статистическое описание классов слов. Статистические методы могут решить эти задачи только в сочетании с данными нестатистической природы. Так, наибольший эффект они дают при описании дистрибуционно-частотных свойств слова. В качестве примера автор приводит один из вариантов такого описания, который может быть использован и для языков тюркской группы.

Доклад А. К. Джубанова (Алма-Ата) «Универсальный алгоритм выделения буквосочетаний тюркских текстов» сопровождался демонстрацией обширного материала, полученного с помощью электронной машины. Предложенные автором алгоритм и машинная программа позволяют получать с помощью ЭВМ алфавитные и частотные списки буквосочетаний от 1 до 24 букв по первым, последним и средним буквам слов тюркских текстов большого объема.

Вопросам информационной нагрузки слов посвящены доклады М. К. Чалабаевой (Ташкент) «Энтропия современного узбекского языка» и Т. И. Ибрагимова (Казань) «Слоговая структура слов татарского языка».

О методике статистического исследования списков сегментов текста фиксированной длины (триад, тетрад и т. д.), которая может быть использована при анализе тюркских языков, сообщалось в докладе Л. И. Белоцерковской (Чимкент) и М. В. Данейко (Минск).

На заключительном пленарном заседании своим опытом по применению статистики и электронных вычислительных машин в изучении структуры языков поделились В. В. Бородин (Горький) и А. И. Бобров (Пермь).

Участники семинара побывали в вычислительном центре и познакомились с возможностями ЭВМ по переработке текстов тюркских и других языков. Им были продемонстрированы все этапы работы машин – от

простейших задач, связанных с получением упорядоченных списков букв, буквосочетаний, словоформ, сочетаний словоформ до автоматического реферирования и перевода.

Участники семинара подчеркнули, что математические и кибернетические методы позволяют вести теоретические исследования в области изучения вероятностно-статистических характеристик тюркских языков на различных языковых уровнях. Семинар отметил определенные успехи в исследовании актуальных вопросов лингвистики и автоматической переработки текстов. Особенно интересной представляется работа, проводимая по статистическому описанию языков классиков казахской литературы Абая и М. Ауэзова.

Намечен ряд конкретных мер по дальнейшему применению статистики и автоматизации в тюркологии. Предполагается регулярное проведение подобных семинаров с изданием их материалов.

Президиум АН Казахской ССР, рассмотрев итоги семинара, обсудил вопрос о дальнейщем развитии автоматизации, лингво-статистических исследований казахского и других тюркских языков. Отмечая важность этих исследований как с точки зрения нормализации современного литературного языка и его истории, так и с точки зрения нужд обработки научно-технической информации, Президиум принял решение укрепить группу «Лингво-статистические исследования и автоматизация» при Институте языкознания АН КазССР.

ИССЛЕДОВАНИЕ СТАТИСТИЧЕСКОЙ СТРУКТУРЫ ТЮРКСКОГО ТЕКСТА

Министерство высшего и среднего специального образования СССР Ленинградское отделение Института языкознания АН СССР. Министерство народного образования МССР Кишиневский политехнический институт им. С.Лазо. Материалы Всесоюзной конференции «Автоматическая переработка текста методами прикладной лингвистики», 6–8 октября 1971 г. Кишинев. – С. 17-18. (Соавт: К.Бектаев, А.Ахабаев)

Группа «Статистико-лингвистического исследования и автоматизация» АН Казахской ССР и Чимкентского педагогического института исследует цепи и комбинаторику лингвистических единиц в тюркском тексте. Составлен и выпущен алфавитно-частотный словарь по произведениям Абая. С помощью ЭВМ получены частотный, алфавитно-частотный и обратный алфавитно-частотный словари эпопеи М.Ауэзова «Путь Абая», в которой на 466 тысяч словоупотреблений приходится около 20 тысяч различных слов и 60 тысяч различных словоформ. В казахских публицистических текстах объемом 150 тысяч словоупотреблений обнаружено 12400 различных слов и 29800 словоформ.

Аналогичные данные получены и по текстам произведений Г.Мусрепова, С.Муканова, Б.Момышулы и по текстам математики и биологии. Статистко-морфологический анализ показывает, что на слова с голой основой приходится 31% словоупотреблений существительных, на падежные окончания – 48%, на окончания притяжательного склонения – 15%, на множественное число – 5%.

При исследовании эпопеи М.Ауэзова эти частоты оказались соответственно разными – 40, 47, 10, 3%.

На зафиксированных в казахской грамматике 500 окончаний существительных практически употребляются лишь 150, которые составляют 98% покрытия (из них 84 дают 98% покрытия), остальные же 350 практически не употребляются.

Одной из отличительных черт тюркских языков по сравнению с индоевропейскими являются их формообразовательная вариативность.

Так, например, в казахском языке 800-900 наиболее частых словоформ покрывают 50% текста.

Исследуется информативная нагрузка тюркского слова. Для этого создается машинная программа автоматического раздробления тюркского словоупотребления на основы и аффиксы.

СТАНОВЛЕНИЕ СТАТИСТИЧЕСКОГО МЕТОДА В КАЗАХСКОМ ЯЗЫКОЗНАНИИ

АН Каз.ССР Институт языкознания.
Труды группы «Статистико-лингвистическое исследование и автоматизация»:
Сб. «Статистика казахского текста».
Алма-Ата, АН Каз. ССР. 1973. С. 14-22.

Казахское языкознание как подлинная лингвистическая наука зародилась только после победы Октябрьской революции. Еще в начале 30-х годов профессор Х.К. Жубанов писал: «Казахское языкознание лишь начинает оформляться в качестве самостоятельной науки. И, как всякая наука, она может завершить стадию своего оформления не на базе общих абстрактных схем, а путем разработки отдельных проблем, чтобы из этих конкретных разработок складывалась наука о казахском языке... Дореволюционное прошлое изучения казахского языка оставило нам весьма скудное наследие, чтобы не сказать «не оставило ничего». Так что почти во всех участках нашей лингвистической работы приходится начинать с азов» [1, стр. 32].

Профессор Х.К. Жубанов родоначальник науки о казахском языке установил наличие всевозможных грамматических категории, выявление фонемно-фонетического состава родной речи, учет терминов казахского языка, созданных и создаваемых на базе лингвистического материала, как

казахского, так и других языков. Большой объем статистических операции, производившихся вручную, был выполнен проф. Х.К.Жубановым и сотрудниками Программно-методического сектора Наркомпроса республики в период создания нового казахского латинизированного алфавита и выработки орфографических правил. Х.К. Жубанов весьма сознательно придерживался принципа количественного подхода к исследованию языке и в связи с освещением порядка построения синтаксических величин, позиционных отношений слов, состава глагольных сочетаний, всевозможных случаев числовых и порядковых отношений грамматических средств и т.д.

Вслед за Х.К.Жубановым в дело систематизации нормативного курса казахского языка включились известные лингвисты: С.К.Кенесбаев, М.Б.Балакаев, Н.Б.Сауранбаев, С.А.Аманжолов и др. Будучи авторами учебников родного языка для начальных и средних казахских школ и педучилищ, они взялись за изучение слоговой структуры казахского языка, физиологии звуков, моментов акцентуации и ударения, случаев прогрессивной и регрессивный ассимиляции и т.д. Для успешного решения этих проблем требовалось выявить количественные характеристики фонологических явлений. Так, в частности, академик АН Каз ССР С.К.Кенесбаев на основе изучения природы казахских гласных и согласных определил количественный состав фонематических единиц, установил структурные типы слогов казахского языка. Все эти положения нашли отражение в учебниках и научных трудах по казахскому языкознанию [2].

Многолетний труд члена корреспондента АН Каз ССР М.Б.Балакаева, посвященный природе синтаксических словосочетаний казахского языка, приобрел большую ценность благодаря количественному анализу типов синтаксических связей. На базе значительного объёма фактического материала ученый выявил наиболее употребительные синтаксические отношения слов типа: глагол+глагол, глагол+имя, имя+глагол, имя+имя и др. По числу компонентов он классифицирует синтаксические сочетании на двучленные (бинарные), трехчленные и так далее. Отдельные типы синтаксических словосочетаний, как установил М.Б.Балакаев, в семантическом отношении порою граничат с устойчивыми сочетаниями слов [3]. М.Б.Балакаев впервые в казахском языкознании высказал мысль, что при установлении лексической нормативности нужно опираться не на субъективную точку зрения, а на частотность употребления лексических единиц, как показатель объективности исходных данных [4,стр.31].

Однако эти попытки использования количественных данных для подтверждения теоретических выкладок носили бессистемный характер, т.к. статическое исследование не носило методический характер. Далее количественные показатели носили бы большую информационную силу только в умелой их лингвистической интерпретации для установления качественных характеристик языковых явлений. Такой подход в рассмотрении лингвистических вопросов не всегда выдерживался.

Более целеустремленно применялся количественный анализ уже в 50-е годы, когда статистический учет грамматических формантов занимал одно из

главных мест в деятельности исследователей языка. Ученые стали больше обращать внимания на документированностъ оперируемых данных. Так, в частности, член корреспондент АН Каз ССР Г.Г.Мусабаев в своей книги о степенях сравнения прилагательных [5, стр. 76-85], не ограничиваясь разрозненными примерами из тексте, дает ряд статических таблиц, иллюстрирующих исследуемые аффиксы по сплошным текстам разных писателей. С этих позиций он подвергает анализу тексты книг: 1) сочинение Абай Кунанбаева 1945 г., 500 страниц; 2) роман «Абай» М.Ауэзова, 300 стр. 3) роман «Жұмбақ жалау» С.Муканова, 1939 г., 500 стр.; 4) роман «Дөң асқан» Г.Сланова, 1939 г., 300 стр.; 5) роман «Өмір мен өлім», Г.Мустафина, 300 стр.; 6) роман «Жас түлектер» А.Абишева, 1945 г., 300 стр.

Значительным успехом казахской лексикографии является издание толково-частотного словаря языка Абая [6]. Работа по составлению словаря была начата большим количеством языковедов под руководством С.К.Кенесбаева. Затем коллективный труд был завершен под рукаводстом профессора А.И.Исхакова и выпущен под его редекциеи в стенах института языкознания Ан Каз ССР. Данной словарь вперые в истории казахской лингвистической науки отразил частоту употребления слов великим казахском поэтом.Составители пользовались методикой построения словария языке Пушкина. Предварительно весь словарный материал произвений поэта был расписанна карточки-цитаты, а затем давалось их толкование. Вся работа выполнялась вручную А.И.Искакова проявил в создании упомянутого

труде большие организаторские способности как лексикограф. Этому способствовали его обширные познания в области лингвистики и опыт практической деятельности, которые сочитались с пониманием ценности нового направления – исследование языка с применением количественного мегода. Исследования, связанные с работой над словарем языка Абая, носили свое освещение в научных трудах ученых. Один из составителей словаря, Т.Жанузаков, в популярной форме анализирует данные толково-частотного словаря [7]. Статистический анализ языка Абая проводится в работах К.Б.Бектаева: «Частотный словарь языка Абая» [8], «Статистика және лингвистика» [9]. Количественный метод использует в своей работе, посвященной синтетическому и аналитическому способам образования глаголов в тюркских языках, профессор В.А.Исенгалиева. Она определила процент употребляемости грамматических средств в процессе создания данной части речи путем снабжения тюркскими аффиксами основ, заимствованных из русского и через русский язык [10]. В результате анализа фактического материала ученый выделяет около шестисот глаголов, подавляющее большинство которых образовано от заимствованных тюркскими языками основ. В конце работы приводятся статистические таблицы с указанием всех грамматических средств, участвующих в образовании глаголе.

В трудах А.К.Хасеновой, посвященных категории глогола казахского языка, такие дается количественная характеристика словообразующих и словоизменительных аффиксов. Ученый дает более или менее полный

перечень залоговых показателей, а также различных глагольных формантов, функционирующих в языке [11].

Переменным моментом в примечании точных (математико-статистических) и инструментальных (с использованием электро-акустической электронной и измерительной аппаратуры) методов в казахском языкознании явилось создание академиком АН Каз ССР С.К.Кенесбаевым лаборатории экспериментальной фонетики и группы Статолингвистические исследования и автоматизация лингвистических работ.

Заведующий фонолаборатории Института языкознания АН Каз ССР Ж.Аралбаев в связи с изучением звукового строя казахского языка, его акустико-артикуляторных характеристик предпринял попытку использования статических методов в определения фонемного состава. Им выявляется вероятность появления фонем в казахском языка, а также информационный вес каждой отдельной фонемы [12]. С.Татубаев впервые в Институте языкознания Ан Каз ССР защитил кандидатскую диссертацию на тему «Статистический анализ спектров гласных» (1968), где предлагается статистических анализ определения характеристик в спектрах гласных, а также способы определения формантных частот и характеристик амплитуда спектров [13]. Сотрудники этой лаборатории М.Раимбекова, А.Джунисбеков, А.Кошкаров, С.Кенесбаев и др., также в значительной степени в своих работах использует количественный метод [14-17].

Современное развитие казахского языкознания предлогает учет не только фонетических, грамматических категорий и морфологических классов слов, но и всего словарного фонда языка.

Ведется большая работа по созданию академического, орфографического, терминологического, диалектологического, этимологического, двуязычного и других типов словарей казахского языка. Безусловно, подобные языковедческие работы предполагают применение статистики.

Ныне Институт языкознания Казахской Академии наук располагает богатым картотечным фондом, который ежегодно пополняется за счет новейших материалов по многим отраслям знаний. Однако такие процессы, как документация миллионов карточек и приведение их в строгий алфавитный порядок, сопряжены о большими трудностями. В этом можно легко убедиться, знакомясь с практикой работы над алфавитно-частотным словарей языка Абая, в создании которого принимали участие десятки научных сотрудников.

В настоящее время при Институте языкознания АН Каз ССР функционирует группа Статистико-лингвистические исследовании и автоматизация лингвистических работ, которой руководит кандидат физико-математических наук, К.Б.Бектаев.

Группа работает в следующих направлениях:

1. Статолингвистическое описание казахского языка (речи) на всех уровнях (фонологическом, лексическом, морфологическом, синтаксическом и т.д.).

2. Определение информационных (кодовых) характеристик языка.

3. Статистическое изучение произведений отдельных писателей.

4. Автоматическая переработка текста.

5. Разработка общих методологических принципов изучений языка и речи в соответствии с современными научными требованиями и т.д. При непосредственном участии К.Б.Бектаев, в также под его руководством члены данной группы: А.Ахабаев, Д.Байтанаев, А.Джубанов, С.Мырзабеков и др. – успешно используют ЭВМ в своих исследованиях при выполнении лексикографических работ в определении энтропии словоформ, статистическом анализе лексико-морфологической структуры современного казахского языка и др. [18-24].

Опыт работы группы показывает, что с помощью новейших достижений техники, в частности электронно-вычислительной машины (ЭВМ), можно добиться большего эффекта в изучении языка и речи, заменить нынешние громоздкие картотеки, создаваемые вручную, быстродействующими и удобными магнитными лентами, перфокартами и т.д., которые во много раз облегчат поиск необходимой языковой информации и обеспечат долгосрочное хранение накопленного материала.

Главное направление деятельности группы определяют решения различных конференций по проблемам статолингвистики: «Межвузовская конференция по вопросам частотных словарей и автоматизации лингвистических работ» (Ленинград, 1966); «Частотные словари и автоматическая переработка лингвистических текстов» (Минск, 1968); «Статистическое и информационное изучение тюркских языков» (Алма-Ата, 1969); «Структурно-математические методы моделирования языка» (Киев 1970); «Статистическое изучение стилей языке и стилей речи» (Горький, 1970); «Автоматическая переработка текста методами прикладной лингвистики» (Кишинев, 1971); «Автоматическое распознавание слуховых образов (АРСО-УП)» (Алма-Ата, 1972).

Следует отметить, что члены группы статолингвистики свои исследования ведут в содружестве с общесоюзной группой «Статистика речи», руководимой профессором Р.Г.Пиотровским [25], активно участвуют во всех названных конференциях и явились организаторами двух Алма-Атинских форумов [26, 27].

Как нас убеждают первые результаты статистического изучения казахского языка (см.[25]) и материалы этого сборника), что вероятностно-статистическое исследование лингвистики с помощью ЭВМ должно получить новые, еще неизвестные науке о языке сведения о количественных отношениях элементов и участков языковой и речевой структур.

Литература

1. Х.Жубанов. Исследование по казахскому языку. Алма-Ата, 1966.

2. І. Кеңесбаев, Ғ.Мұсабаев. Қазіргі қазақ тілі. Алматы, 1962.

3. М.Балакаев. Современный казахский язык. Изд-во «Наука». Алма-Ата, 1959.

4. М.Балақаев. Қазақ тілі мәдениетінің мәселелері. Алматы, 1965

5. Ғ.Мұсабаев. Қазақ тіліндегі сын есімнің шырайлары. Алматы, 1951.

6. Абай тілінің сөздігі. Алматы, 1970.

7. Т.Жанұзақов. Тіл біліміндегі статистика. «Лениншіл жас», 1965, 7 июль, № 133.

8. К.Б.Бектаев. Частотный словарь языка Абая. МКЧС, Л.,1966.

9. Қ.Б.Бектаев. Статистика және лингвистика. Журнал. «Білім және Еңбек». Алматы, 1967, №6.

10. В.А.Исенгалиева. Тюркские глаголы о основами, заимствованными из русского языка. Алма-Ата, 1966.

11. А.Қ.Хасенова. Етістіктің лексика-грамматикалық сипаты. Алматы, 1971.

12. Ж.А.Аралбаев. Вокализм казахского языка. Алма-Ата, 1970.

13. С.Татубаев. Статистический анализ спектров гласных. АКД, Алма-Ата, 1968.

14. М.Раимбекова. Длительность гласных в казахском языке. АКД, Алма-Ата, 1968.

15. А.Джунусбеков. Гласные казахского языка. АКД, Алма-Ата, 1969.

16. А.Кошкаров. Спектральный анализ фрикативных согласных казахского языка в структурах типа СГС, ГСГ. АКД, М., 1972.

17. С.С.Кенесбаева. К вопросу исследования протетического звука в словах арабского происхождения статистическим методом. Сб, «Статистическое и информационное изучение тюркских языков». Алма-Ата, 1969.

18. С.К.Кенесбаев, К.Б.Бектаев, Р.Г.Пиотровский. Инженерная лингвистика и тюркология. «Советская тюркология», Баку 1970, №6.

19. С.К.Кенесбаев, К.Б.Бектаев, Р.Г.Пиотровский. О статистическом изучений тюркских текстов. СИИТЯ, Алма-Ата, 1969.

20. К.Б.Бектаев, С.К.Кенесбаев, Р.Г.Пиотровский. Инженерная лингвистика. «Вопросы языкознания» 1973, № 2,

21 С.К.Кенесбаев, А.Т.Кайдаров. Казахское языкознание за 50 лет. «Вопросы языкознания, 1973, № 1.

22. В.Л.Лобин. О принципе акустического и фонологического отбора в развитии языка. АКД. Алма-Ата, 1969.

23. А.Ахабаев. Статистический анализ лексико-морфологической структуры языка казахской публицистики (на материале газетных текстов). АКД. Алма-Ата. 1971.

24. Қ.Б.Бектаев. Қазақ тілінің кері алфавитті сөздігі. Алматы 1970.

25. К.Б.Бектаев. Статистика речи 1957-1972 г.г. (Библиографический указатель). Алма-Ата, 1972.

26. Статистическое и информационное изучение тюркских языков. Тезисы докладов и сообщений. Алма-Ата, 1969.

27. Труды VII Всесоюзной школы-семинара «Автоматическое распознавание слуховых образов». Алма-Ата, 1973.

СТАТИСТИКА ЛЕКСИКО-МОРФОЛОГИЧЕСКИХ ФОРМ ПРИЛАГАТЕЛЬНЫХ В РОМАНЕ М. АУЭЗОВА «АБАЙ ЖОЛЫ» («ПУТЬ АБАЯ»)

АН Каз.ССР Институт языкознания.
Труды группы «Статистико-лингвистическое исследование и автоматизация»: «Статистика казахского текста». Алма-Ата, «Наука». 1973. С. 231-248. (соавтор: Е.Х. Жубанов)

Цель данной работы заключается в статистическом изучении лексико-морфологической структуры прилагательных казахского языка на материале романа «Абай жолы» с частичным сопоставлением с материалами публицистики.

Как известно, имена прилагательные казахского языка обладают целым рядом признаков, отличающих их от остальных частей речи. Во-первых, исконно казахские прилагательные представляют собой корневые морфемы. Во-вторых, прилагательные казахского языка обладают присущими только им морфологическими признаками, указывающими на степени сравнения. На основе именно этого признака многие тюркологи выделяют прилагательные как самостоятельную часть речи. Третья особенность прилагательных тюркских языков, в том числе и казахского, заключается в том, что эта категория слов в определенных случаях снабжается в начальной позиции усилительным слогов или так называемым интенсивом. Четвертым признаком прилагательных следует признать их семантическую устойчивость. Характерной особенностью их, например, является то, что они, выступая в роли наречия, не теряют своего значения. Все эти специфические формы и семантические особенности имени прилагательного казахского языка четко отграничивают его от других частей речи как особое лексико-грамматическое образование. По своему значению прилагательные казахского языка делятся на две группы: качественные и относительные.

В исследуемом нами тексте романа имена прилагательные использованы писателем 38531 раз, что составляет 8,3% от всех словоупотреблений (465966). Без учета повторяемости различные словоформы прилагательных встречаются 4773 раза. Это 7,7% всех словоформ в романе. Процентные соотношения довольно ясно оттеняют стиль писателя, который равномерно употребляет слова-прилагательные в тексте романа.

Классификация прилагательных производилась по семантико-морфологическим признакам. В приводимой нами таблице представлены данные о формах прилагательных и указано их процентное соотношение, а

также процент покрываемости форм прилагательных вычислен по их отношению к двум данным: а) общее количество различных форм прилагательных (длина словника), б) общий объём прилагательных в тексте (с учетом повторяемости).

Писатель использовал все существующие формы казахских прилагательных, дополняя их даже некоторыми русскими прилагательными типа: крестьянский, штатский, польский, становой, мировой, окружной, волостной, крепостной, уездный, тайный, государственный, военно-полевая, русско-киргизская, учительская (семинария) и т.д. Как видно, все эти русские прилагательные включены в роман как историзмы или архаизмы, характеризующие описываемую эпоху. Разумеется, они не входят в словарный фонд современного казахского языка, а только находят применение в художественном тексте. Общее количество таких заимствованных слов не составляет и 1% всех прилагательных. Что касается повторяемости их в тексте, то она и того меньше (0,2%).

Таблица

Формы прилагательных	Процент от длины словника прилагательных	Процент от всего объема прилагательных в романе
Корневые	14,9	47,7
Аффиксальные	66,4	45,4
Из парных слов	5,5	1,6
Из слитных слов	5,4	1,8
Степени сравнения	5,1	2,7
Интенсивные	1,2	0,6
С нетюркскими аффиксами	0,7	0,4
Заимствованные	0,8	0,2
Всего прилагательных	4773 (100%)	38531 (100%)

В таблице учитываются и прилагательные, образованные от казахских слов, с прибавлением к ним нетюркских (иранских) аффиксов типа: *алда-м-паз, боқта-м-паз, әсем-паз, өнер-паз, ал-ым-паз, шал-ым-паз, жаг-ым-паз, ресм-и, әскер-и, дүние-үй, щарх-и, бей-қам, бей-маза, би-шара, би-тарап, бәд-бахыт, на-разы, на-риза* и т.д.

Немногочисленные по своему составу, эти аффиксы давно вошли в язык казахов и воспринимаются ныне как исконно казахские элементы речи.

В романе нередко встречаются и некоторые слова иноструктурного типа, которые использованы писателем для создания образов отдельных персонажей романа, как, например: *ақли, ғибади, тахлиди, шағи*. Заимствованные из арабо-иранских языков, эти слова религиозного обихода не прижились в казахском языке. Что касается заимствованных прилагательных *ғазиз* и *фак*, а также их фонетических вариантов *ғазіз, әзіз* и *пак, пәк* встречающихся в романе, то они рассматриваются как синонимы аналогичных казахских прилагательных. Прилагательные, образованные путем прибавления нетюркских аффиксов, составляют незначительный процент как в словнике, так и в самом тексте. Но заметим, что по частоте

встречаемости в романе они вдвое превосходят употребленные автором русские прилагательные.

Особую форму среди казахских прилагательных представляют лексико-морфологические образования с усилительным слогом в препозиции. Интенсивная форма прилагательных – это те же самые имена прилагательные, которые для большей экспрессии наращивают впереди себя созвучный слог. В романе такая форма прилагательных составляет 1,2% словника, а в целом его частота 0,6% от общего числа прилагательных всего текста. В речи писателя с наибольшей частотой употреблены прилагательные *қап-қара* (30), *жым-жырт* (25), *оп-оңай* (20), *сүп-сұр* (14), *құп-қу* (15), *жіп-жіңішке* (13). Частота остальных слов менее десяти. По способу образования эти прилагательные не представляют ничего особо индивидуализированного, отличающегося от общенародной лексики.

Особый интерес в стиле писателя представляет употребление им степеней сравнения прилагательных. Как известно, прилагательные только при образовании степеней сравнения могут принимать аффиксальные элементы. Так сравнительная и уменьшительно-приблизительная степени сравнения прилагательных имеют свои формообразующие аффиксы: -рақ//-рек, а также *-лау//-леу, -дау//-деу, -тау//-теу, -ғылт//-ғылтым, -ғыш, -шым//-шіл* и т.д.

Исключение составляет преувеличительная степень, которую представляет упомянутая выше интенсивная форма, и превосходная степень, для образования которой необходимо сочетание имени прилагательного с усилительными словами типа: *ең, тым, аса, өте, тіпті* и т.д. – со значением «очень», «весьма», как, например: *ең жақсы* (самый хороший), *тым күшті* (очень сильный), *аса бағалы* (весьма ценный) и т.п.

Статистика формообразующих аффиксов показывает, что различные формы степеней сравнения, образованные аффиксальным путем, составляют 5,1% всех остальных прилагательных, а с учетом повторяемости – 2,71% от общего числа прилагательных, испольвованных в романе.

Известный лингвистический интерес представляют принципы подхода писателя к формообразующим аффиксам прилагательных. Так аффикс *-қай* употребляется автором всего 45 раз. Из них в трех случаях он прибавляется к трем прилагательным: *шым-қай, сұр-қай, тырбы-қай*, а в остальных 42 случаях снабжает только прилагательное *қоңыр* (темный) в виде *қоңырқай*. Такое явление объясняется тем, что слово *қоңырқай* многозначное и может служить для обозначения признаков не только предметов, но и различных явлений природы и общества. Среди цветовых оттенков наибольшую частоту имеет слово *қызғылт* (красноватый) – 55, далее следуют приведенное выше *қоңырқай* – 45, *көкшіл* (синеватый) – 41, *сұрғылт* (сероватый) – 38, *ақшыл* (беловатый) – 37, *сарғыш* (желтоватый) – 30, остальные прилагательные, выражающие оттенки цветов, встречаются менее часто, как, например: *бозғыл* – 9, *сұрқыл* – 3, *қуқыл* – 3, *шымқай* – 1, *сұрқай* – 1. Среди этих слов, обозначающих цветовые оттенки, привлекают внимание такие синонимические слова, как *сұрғылт, сұрқыл, сұрқай*, образованные от

прилагательного *сұр* (серый). Но наиболее типичным из них является *сұрғылт*, ибо аффикс *-ғылт* в казахском литературном языке оформляет только это слово. Что касается аффиксов *–қай* и *–қыл*, то они обычно прибавляются не к *сұр*, а к другим прилагательным типа: *қоңыр, боз,* қу, *шым* и т.д. Ясно, что прилагательное *сұрқай* является неологизмом, созданным автором путем необычной грамматикализации и подчеркивающими особенность стиля писателя.

Из числа других прилагательных со степенями сравнения можно упомянуть образование *қатқыл* (31). В романе этим словом обозначается признак тона говорящего и употребляется более часто, чем его синонимы *қатаң, қатал, қатты* и т.д. Заслуживает также особого упоминания прилагательное *жүдең*. Хотя это слово встречается в романе один раз, необходимо признать его фактом индивидуального словотворчестве автора. Литературный вариант этого прилагательного – *жүдеу* выступает в языке как имя действия, образованное он глагола *жүде.* Однако писатель оформляет данное слово по типу прилагательных *ашаң, жұтаң, қуаң* и т.д. со значением истощенный, худощавый, захудалый и др. Этим самым автор вскрывает не обнаруженную ранее потенциальную возможность лексической единицы.

Процентные соотношения прилагательных, имеющих формы сложных и слитных слов, примерно равны. Это и понятно, ибо подобные структуры занимают значительный объём в составе морфологических классов слова и представляют собой активные способы формообразования в казахском языке. Так в словнике обе словоформы прилагательных представлены почти одинаково: прилагательные из сложных слов составляют 5,5%, а из слитных слов – 5,4% всех форм. Соответственно их частоты употребления выражены в процентах: 1,6% и 1,8%, что также не содержит больших отклонений.

В приведенной выше таблице процентных соотношений различных форм прилагательных нами зафиксирована основная масса этой части речи как корневые и аффиксальные (производные) формы прилагательных. По своей лексико-грамматической сути эти формы соответственно обозначают качественные и относительные прилагательные. Наиболее общей чертой этих двух форм, судя по их статистике, можно назвать примерно одинаковую частоту употребления их в тексте. Так, если корневая форма прилагательных составляет 47,7% всех словоупотреблений прилагательных, то аффиксальная форма – 45,4%. Однако такое равновесие в употреблении прилагательных исходит не из тождества их словника, т.е. использованных словоформ. Наоборот, в авторском употреблении словоформы качественных прилагательных, т.е. корневых, составляют 14,9%, словника, тогда как производным формам, т.е. относительным прилагательным, отводя 66,4%. Это означает, что писатель в целом по роману больше использовал качественные прилагательные, выражающие признаки предметов и явлений в своей морфемной структуре, нежели относительные прилагательные, которые могут определить то или иное свойство описываемого предмета, только исходя из характеристик аналогичных явлений.

Среди качественных прилагательных, выражающих различные цвета и масти животных, с наибольшей частотой использовано прилагательное *қара* (черный) – 644 раза. С высокой частотой употребляется и его антоним ***ақ***, (белый) – 406 раз. По-видимому, это объясняется тем, что слова *черный* и *белый* в казахском языке обозначают не только внешнюю окраску предмета, но и такие абстрактные понятия, как виновный и невиновный, злой и добрый и т.д.

Из остальных слов, обозначающих цвета, более часто употреблены слова: *көк* (синий) – 226, *қызыл* (красный) – 200, *қоңыр* (коричневый) – 171, *сары* (желтый) – 145. Более 50 раз употреблены в романе слова *сұр*, *бурыл, жирен*. Кроме них в художественной палитре писателя имеются прилагательные: *күрең, мөлдір, жасыл* и т.д.

Не менее интересен подбор прилагательных писателем для характеристики психо-физиологических свойств. Примечательно, что в этом разряде самая большая частота принадлежит прилагательному *қатал*, (суровый) – 102. Слово *момын* (тихий, безвредный) втречается 96 раз. За ним следует прилагательное *сыпайы* (вежливый) – 92, *адал* (честный) – 78, *нәзік* (нежный) – 69 раз. С одинаковой частотой употреблены прилагательные, обозначающие положительные окраски: *әділ* (справедливый) – 45, *сергек* (чуткий) – 45, *батыл* (смелый) – 44. Равномерно распределяется также целый ряд прилагательных с отрицательной окраской. К ним можно отнести слова: *қас* (враждебный) – 21, *сотқар* (задиристый) – 27, *содыр* (забияка) – 21, *арам* (зловредный) – 28, *баяу* (медлительный) – 29. Из прилагательных подобного типа сравнительно большую частоту имеют слова: *қу* (хитрый) – 66, *тентек* (буйный) – 65 и т.д

Антонимия синомических рядов качественных прилагательных с частотой выше 30-ти

1.	жақсы – 652, сұлу – 215, әсем – 76 асыл – 87, таза – 89, ырза – 55, сау – 40, жүйрік –59, шебер – 47, лайық – 31, сай – 28, тең – 42	1.	жаман – 191, надан – 40, ащы – 44
2.	бай – 31, мол – 246, кең – 308	2.	кедей – 44
3.	семіз – 86	3.	жүдеу – 70
4.	Қатты – 904, берік – 51, мықты – 69	4.	Бос – 40, жай – 71, еркін – 51
5.	жаңа – 230	5.	ескі – 64
6.	жас – 565	6.	кәрі – 159, қарт – 53
7.	оңай – 96	7.	қиын – 87
8.	жеңіл – 91	8.	ауыр – 267
9.	Үлкен – 1200, ірі – 41, зор – 166, кесек – 83, тұтас – 31, қалың – 78, жуан – 104, биік – 167, терең – 78	9.	кіші – 65, кішкене – 177, ұсақ – 76, кем – 30, шолақ – 47, орта – 89, аласа – 59
10.	шын – 206, рас – 103, анық – 319, ашық – 159	10.	жалған – 53, құпия – 31
11.	ыстық – 204, жылы – 93	11.	суық – 225, салқын – 170
12.	алыс – 98, ар – 157, жат – 60, бөтен – 56, сырт – 46, өзге – 360, бөгде – 63,	12.	жақын – 282, бер – 30, таныс – 49, қымбат – 52

	басқа – 245, бөлек – 155		
13.	оң – 110	13.	теріс – 57
14.	ұзын – 207, ұзақ – 142	14.	қысқа – 118

Качественные прилагательные, выражающие различные свойства предметов и явлений, семантически переплетаются двумя способами: как антонимы и синонимы. Причем синонимические ряды вступают между собой в антонимические отношения. Разумеется, по частоте употребления подобные противопоставления не могут соблюдать какого-либо равновесия, мало того, в зависимости от стилистического употребления автора они нередко переходят из одного разряда в другой. Однако, оставляя право исследования подобных лексико-семантических аномалий за лингвистами-традиционалистами, здесь мы приводим для наглядности отношения синонимических рядов с точки зрения их антонимии. В данный перечень вошли прилагательные с частотой более тридцати, которые отражают качественные состояния предметов.

Для более ясного представления статистики аффиксов прилагательных мы сопоставили данные II книги романа «Абай жолы» (124398 словоупотреблений) с публицистическим текстом современного казахского языка (150000 словоупотреблений).

По нашим подсчетам, в указанной книге романе встречается 11467 (9,2%) прилагательных с учетом повторяемости. Аналогичный объём публицистического текста содержит 18534 (12,4%) прилагательных. Из всего числа прилагательных, использованных М.Ауэзовым, корневые слова составляют 51,3%, а остальные, т.е. образованные аффиксальным путем 48,7%. Судя по этим отношениям, можно утверждать, что в романе соблюдено приблизительное равенство между этими двумя видами имен прилагательных. В отличие от романа в публицистике аффиксальная форма прилагательных заметно превалирует над корневыми, т.е. прилагательные, образованные путем прибавления аффиксов, составляет 70%, тогда как корневые покрывают 30% всех прилагательных. Отклонение между двумя видами прилагательных достигает более 40% против 2% в исследуемом романе М.Ауэзова. Такое различие между двумя жанрами, по-видимому, объясняется тем что художественному тексту свойственна большая экспрессивность, тогда как газетные и журнальные тексты используют более сдержанные выражения в смысле красочности. Это и способствовало тому, что публицистике преобладают производные прилагательные, обозначающие отношения к месту, времени и т.п.

Что касается статистики только аффиксов (как в романе, так и в публицистике), то их подсчет осуществлялся нами по трем разновидностям: аффиксы отыменных прилагательных, отглагольных прилагательных и малопродуктивные древние аффиксы.

Как было указано выше, аффиксальные прилагательные публицистического текста доминируют над корневыми и на 21% превышают

аналогичные прилагательные из текста «Абай жолы». Однако, несмотря на эту разницу, общее количество аффиксов каждого текста почти одинаковое.

Так, если в тексте публицистики аффиксы отыменных прилагательных употреблены на 1,4% больше, чем в тексте романа, то, наоборот, аффиксов отглагольных имен прилагательных в публицистике на 0,7% меньше. Процентные соотношения употребления малопродуктивных и древних аффиксов а обоих текстах также не имеют больших отклонений (0,7%). Если взять более конкретные данные, то общая картина соотношения аффиксов станет более ясной. Так, в романе 38 формантов отыменных прилагательных употреблены 4987 раз, 21 формант отглагольных прилагательных употреблен 365 раз, 22 древних аффиксов – 235 раз. Соответственно в текстах публицистики имеются 39 формантов отыменных прилагательных, которые повторяются 11832 раза, 27 формантов отглагольных прилагательных о частотой 768 раз и 25 древних аффиксов, которые употреблены 375 раз.

Частотное распределение классов относительных прилагательных

классы аффиксальных прилагательных	**роман**				**публицистика**			
	Колич. аффиксов	**%**	**Частота аффик-сов**	**%**	**Колич. аффик-сов**	**%**	**Частота аффик-сов**	**%**
Именной аффикс	38	46,9	4987	89,8	39	42,9	11832	91,2
Глаголный аффикс	21	25,9	365	6,6	27	29,71	768	5,9
Древний аффикс	22	27,2	235	3,6	25	27,4	375	2,9
Итого	81	100	5587	100	91	100	12975	100

В процентном отношении нет больших расхождений в употреблении аффиксов отыменных основ. Однако не все аффиксы в одинаковой степени употреблены как в романе, так и в публицистике.

Равномерность распределения аффиксов отсутствует и внутри отдельных текстов. Так в романе «Абай жолы» (II кн.) 12 аффиксов из 38 составляют 88% употребляемости. Только один формант ***-ғы*** использован писателем с частотой 1006 раз, что составляет 20,2% всей частоты. Не без интересно также проследить распределение мягких и твердых вариантов в тексте. Если мягкие варианты указанных аффиксов составляют 34,8%, то твердые варианты – 65,2%. Причем 11% из 34,8% принадлежит только форманту – ***гі***.

Аналогично в тексте публицистики 82,2% употребляемости падает на 12 аффиксов из 39. Такие аффиксы, как ***-ғы//-гі,-сыз//сіз, ды//-ді, -ты//-ті,-лы//-лі, -дай/-дей***, в романе и в публицистике находятся в зоне высокой частоты. В отличие от них аффиксы: ***-дық//-дік, -тық//-тік, -лық//-лік***,

имеющие высокую активность в публицистике, не являются таковыми в романе.

Эти аффиксы не могли быть активными в романе потому, что писатель отражал в нем события прошлого столетия из жизни казахов. Хотя формант ***лық***, и др. его фонетические варианты являются исконно казахскими, в досоветский период они употреблялись только со считанными словами типа: *адам+шы+лық, бала+лық, аңшы-лық, жас+тық, болыс-тық, кәрі+лік, ит+тік* и т.д., которые выражали отношение и тому или иному предмету в жизни общества. В современный период развития языка, когда сфера функционального развития казахского языка немного расширилась, данный аффикс получил наибольшую активность в качестве словообразовательного форманта, который может прибавляться не только к исконно казахским, но и также к заимствованным основам, как например: *социалистік*, *участоктық, демократтық, революциялық* ранее не бытовавшим в языке. В тексте публицистики твердые и мягкие варианты аффиксов встретились почти в одинаковой степени. Так, если процент твердых аффиксов 51,6% то мягких – 48,4%. Частотные характеристики всех отыменных аффиксов даны в таблице.

Частотные характеристики аффиксов отыменных прилагательных

№ п/п	Аффиксы	Роман	Публицис-тика	№ п/п	Аффиксы	Роман	Публицис-тика
1	***-ғы***	1006	2022	22	***-лік***	19	180
2	***-гі***	547	1248	23	***-дық***	16	602
3	***-қы***	77	273	24	***-дік***	14	180
4	***-кі***	64	92	25	***-тық***	20	226
5	***-ты***	561	674	26	***-тік***	3	852
6	***-ті***	138	664	27	***-сыз***	196	254
7	***-ды***	171	421	28	***-сіз***	262	149
8	***-ді***	189	199	29	***-шыл***	60	82
9	***-лы***	389	414	30	***-шіл***	25	65
10	***-лі***	193	457	31	***-қор***	9	8
11	***-дай***	523	450	32	***-қой***	6	8
12	***-дей***	203	96	33	***-и***	5	421
13	***-тай***	87	40	34	***-ы***	15	-
14	***-тей***	30	7	35	***-і***	11	5
15	***-лас***	33	17	36	***-паз***	5	-
16	***-лес***	20	62	37	***-шаң***	-	1
17	***-дас***	8	14	38	***-шең***	-	1
18	***-дес***	9	5	39	***-ный***	-	6
19	***-тас***	1	23	40	***-ғыдай***	2	89
20	***-тес***	5	10	41	***-гідей***	2	17
21	***-лық***	63	1498	Итого: 4987			11832

Отглагольные прилагательные романа «Абай жолы» образованы при помощи различных формантов. Это составляет 25,9% всех аффиксальных имен прилагательных. Они встречаются в тексте с частотой 365 раз, что есть 6,6% от всей частотности прилагательных. Аналогичные показатели публицистического текста имеют некоторые расхождения в процентном

отношении. Так здесь 27 различных форматов составляют 29,7% от общего числа аффиксов, повторяемость которых равна 766, т.е. 5,9% всей накопленной частоты.

Хотя в количественном отношении аффиксы прилагательных встречаются в равной мере как в романе, так и в публицистике, в отношении самих аффиксов существует некоторый разнобой. Скажем, в романе отсутствуют форманты *-ңқы//-ыңқы, -ба//-бе//-пе, -палы, -інді, -шек, -ақ/-кіш*, хотя, за исключением форманта *-палы*, все они по частоте ниже 10 употреблений. В свою очередь в текстах публицистики отсутствуют аффиксы *-мды//-ымды, -нды*, которые в романе играют заметную роль в смысле частоты употребления. Так из этих формантов только *-мды* и *-ымды* употреблены менее 10 раз. Аффиксы *-іңкі, -мды, -іс* имеют частоту от 10 до 20. Что касается форманта -ыс, то он использован во II книге романа более чем 50 раз, тогда как в публицистическом тексте зарегистрирован всего один раз. В процентном отношении эти два аффикса *-с, -ыс* – покрывают 30,4% аффиксального образования. В научной грамматике казахского языка форманта *-с, -ыс//-іс* отнесены к разряду непродуктивных аффиксов, хотя данные романа отражают иную картину. Писатель образовывает целый ряд прилагательных типа: *тозбас, аралас, қимас, қажымас, айнымас, жасымас, аямас* и т.д., которые хотя и с малой частотой, но встречаются довольно активно, снабжая множество глагольных основ. Так, наибольшая частота из всего числа (56) прилагательных, образованных при помощи ***-с***, принадлежит словам: *таныс*-7, *аралас*-6, *қимас*-5. Среди 55 прилагательных, оформленных аффиксом ***-ыс*** относительно часто встречается слово ***туыс*** (3). Это означает, что писатель, владея богатым словарным фондом, не прибегает к частому употреблению отдельных словесных величин, хотя и многократно использует известные форманты с различными корневыми морфемами.

Примечательно, что в романе из 30 разновидностей формантов, образующих отглагольные прилагательные, только шесть, именно: *-ы// -і, -імді, -с// -ыс, -лы* – составляют 70,7% всего употребления, остальные 24 форманта *-мды, -ғыш//-гіш, -қыш//кіш, -ба//-бе, -па//-пе, -ма, -малы, -палы//-пелі, -нды, -ынды//-інді, -ңқы, -ыңқы//-іңкі, -шақ//-шек, -ақ, -лі* покрывают лишь 29,3%.

Подобно этому наиболее продуктивные 8 аффиксов в языке публицистики покрывают 90,1% всей накопленной частоты отглагольных прилагательных. На долю остальных 19 формантов остается 9,9%. В число этих 8 аффиксов входят форманты: *-с, -імді, -ғыш//-гіш, -ма, -малы, -лы//-лі*. Из этих разновидностей аффиксов наибольший процент употребляемости принадлежит форманту ***-ма***, который встречается 287 раз, что составляет 37,25% всех остальных форм. При этом формант ***-ма*** в тексте публицистики оформляет лишь два слова – *құра-ма* (273 раза) и *тоқы-ма* (14).

Если исходить из твердости и мягкости аффиксов, то их статистика выглядит следующим образом. Твердые варианты аффиксов отглагольных прилагательных составляют в тексте романа 57,5%, в публицистике - *68%*. Однако здесь следует учитывать высокую частотность форманта ***-ма*** в

последнем тексте, ибо без него процент твердых аффиксов равнялся бы 30,8%. Мягкие варианты аффиксов соответственно составляют 42,5% в романе и *32%* в публицистике.

На примере этих данных можно проследить наибольшую последовательность сингармонизации слов у писателя, нежели в языке публицистики. Равномерно распределяя в тексте твердые и мягкие варианты фонематических структур, М.Ауэзов достигает благозвучия и музыкальности ритмизированной речи, и тем самым эвфония его словесной ткани становится сопричастным элементом художественного речетворчества. С другой стороны, подобная черта романа подчеркивает обязательность внутренней гармонии для эстетического целого, каковым является произведение литературы, чего однако нельзя требовать от разрозненных публицистических текстов.

В научной грамматике казахского языка учтено среди словообразующих аффиксов прилагательных более 30 формантов, так называемых древних или, как еще принято обозначать их в научном обиходе, непродуктивных аффиксов. Эти форманты, участвуя в словообразовании в ограниченной сфере, встречаются в современном языке, как бы закрепленными за отдельными корневыми морфемами. Частота подобных аффиксов, естественно, бывает невысокой. В нашем исследовании установлено, что во II книге романа «Абай жолы» из числа древних аффиксов употреблены 22 формы. Что касается выборки из публицистики, то в них немного больше, т.е. наличествует 25 таких аффиксов. С точки зрения частотных характеристик, обнаружено, что реализация в речи древних формантов имеет свои особенности в сравниваемых текстах.

В тексте романа из 22 выявленных нами формантов только 7 составляют 85,5% от всей употребляемости. Это аффиксы: ***-кір*** (17), ***-ық*** (123), ***-есі*** (-11), ***-гек*** (12), ***-іметал*** (12), ***-ғақ*** (14), ***-қақ*** (12). Они представляют собой зону высокой частоты (частота более 10) в среде древних аффиксов ауэзовского употребления.

В языке же публицистики 7 аффиксов с частотой 10 и больше покрывают 78,9% в среде 25 формантов, повторяющихся 375 раз. К этим пяти аффиксам относятся: ***-ық*** (206), ***-ік*** (21), ***-ғысыз*** (11), ***-есі*** (25), ***-кер*** (10), ***-мпаз*** (10), ***-келең*** (13). Однако надо учесть то, что формант ***-ық*** из общего числа повторяемости только 123 раза прибавляется к одному и тому же корню ***тол+ық***, что есть 32,8% из всей накопленной частоты древних аффиксов.

Как видно из этих перечислений, не одни и те же аффиксы образовали высокочастотную зону в сравниваемых нами двух текстах. Общими для них оказались только лишь форманты ***-ық, -есі,*** которые одинаково попали в высокочастотную зону. Вообще при сопоставлении употребляемости в этих двух текстах установлено, что только 16 формантов ***(-қыр, -кір, -гер, -ық, -ік, -іметал, -ғыр, -гір, -дақ, -ғақ, -қақ, -есі, -ымпаз, -ғысыз, -гісіз, -тар)*** употреблены как в романе, так и в публицистике. Остальные же аффиксы в романе: ***-ымтал, -дек, -ымсақ, -тек, -гілікті, -кілікті*** и в публицистике: -

кер, -ыр, -сақ, -дар, -келең, -мпаз, -імпаз, -қысыз, -кісіз, -машы взаимно исключают друг друга в двух различных жанрах.

Эти факты способствуют установлению стилистической дифференциации между языками художественно-индивидуализированного произведения и публицистики. Такова общая картина статистического анализа при сопоставительном исследовании двух стилей.

ИНДЕКСАЦИЯ И КОДИРОВАНИЕ ТЕКСТОВОЙ ИНФОРМАЦИИ ДЛЯ ВВОДА В ЭВМ

АН Каз.ССР Институт языкознания.
Труды группы «Статистико-лингвистическое исследование и автоматизация»: «Статистика казахского текста». Алма-Ата, «Наука». 1973. С. 250-262. (соавтор – К. Б. Бектаев)

Лингвистический аспект подготовки текста. Наиболее кропотливой работой, выполняемой вручную в процессе автоматического анализа лингвистических единиц, является предмашинное редактирование текста. При этом интересующие нас словесные единицы снабжаются условными пометами – индексами. Такая необходимость возникает из-за ограниченной, по сравнению с человеком, эвристической способности современной ЭВМ, ибо анализируемые лингвистические единицы текста не всегда могут быть полностью распознаваемыми автоматически точно на основе их формальных признаков. А составление программы, с помощью которой ЭВМ могла бы определить принадлежность, например, каждого словоупотребления к тому или иному грамматическому классу, представляет большую сложность, поэтому удобней всего снабдить слова и словоформы индексами, указывающими на их грамматический класс. Эта дает возможность осуществить статистическое описание на уровне конфигурации лексических единиц, а также устранить лексико-грамматическую неоднозначность (многозначность и омонимию единиц текста), являющаяся основной лингвистической трудностью при автоматическом анализе текста.

Как и во многих других языках, значительное количество казахских слов составляет внутри языка омонимические ряды, и потому их принадлежность к той или иной категории частей речи выявляется только в контексте. Например, казахские слова «***қара***» может быть глаголом (смотри), прилагательным (черный), существительным («скот» перен.).

Следовательно, подобные единицы языка мы заранее отвечаем в тексте условными индексами, которые и будут учитываться при дальнейшей подготовке текста к вводу в ЭВМ и при составлении необходимых программ. Первоначально в качестве опыта нами была выработана условная система индексов, состоящих из следующих девяти букв: З – существительное, Г – глагол, С – прилагательное, У – наречие, А – числительное, Е – местоимение,

О – междометие, К – служебное слово, Л – подражательное слово, Д – модальное слово. В настоящее время применяется следующая индексация: 3 (цифра) – существительное, 7 – глагол, 6 – прилагательное, Ю – наречие, А – числительное, 4 – местоимение, О – междометие, Ш – служебное слово, Л – подражательное слово, 5 – модальное слово (әр тарап, бейтарап сөздер), 1 – служебные имена, 2 – имена собственные (человека), Ж – название животных, Х – сокращенное слово, Н – название культурно-просветительных, бытовых учреждений, Щ – название научных трудов, учебников, произведений, Э – географическое название, Т – название государств, племени рода.

В дальнейшем, в связи с расширением круга задач, выяснилось, что указанные индексы не полностью удовлетворяют наши запросы, ибо наряду с грамматическими признаками слова надо было отметить и их лексико-статистические свойства. С учетом всего этого впоследствии нами была разработана новая, буквенно-цифровая система условных индексов, указывающих на синонимические, омонимические, паронимические и антонимические ряды в составе того или иного текста (преимущественно художественного). К тому же с целью получения списков лексических пластов типа архаизмов, неологизмов, модернизмов и т.д. разработаны дополнительные индексы.

В традиционном языкознании слова классифицируются на основе трех главных признаков: лексического значения, морфологической структуры и синтаксической функции. Совокупность этих трех признаков дает право отнести то или иное слово к определенному разряду частей речи на уровне морфологии. Следовательно, индексирование слов по частям речи являет собой главную необходимость при статистическом изучении.

В принципе наша индексация носит фронтальный характер, т.е. помечаются все словесные величины в составе изучаемого текста. Однако это не исключает возможности индексировать в тексте только выборочные слова или сочетания слов, в зависимости от поставленной цели.

Кроме перечисленных выше индексов в практике машинного анализа применяются также различные знаки для размежевания основы и окончания, корней слитных слов, слогов, для обозначения признака сочетаемости лексических единиц. Все эти делиматры являются формальными диагностическими признаками (ФДП). В свою очередь, делиматры различаются как абсолютные и относительные. Проставляемые нами буквенные, цифровые и др. служебные символы суть относительные делиматры. Что касается тех формальных признаков, которые имеют место в самом тексте, как например, пробелы, знаки препинания, тире и дефисы, то они воспринимаются как абсолютные делиматры. Для выделения из текста словоупотреблений достаточно, например, код пробела, а для предложений – код точки, т.е. абсолютные делиматры. Введение в текст относительных делиматров, как например: =, + , / , X, < , > и т.д. дает возможность программисту выделить устойчивые словосочетания, фразеологизмы, терминологические группы и т.п. Желательно редактирование

(индексирование) текста производить при решении максимального количества лингвистических задач, т.к. дальнейший этап подготовки текста для ввода в память ЭВМ является довольно трудоемким.

Инженерный аспект подготовки текста. После полного редактирования лингвистического материала начинается этап подготовки текста к восприятию (или «чтению») ЭВМ. Для чего необходимо преобразовать текст из графемной (или звуковой) формы в особую «машинную» форму таким образом, чтобы тождественность и нетождественность отдельных единиц текста не были нарушены. Этот процесс осуществляется при участии человека-оператора и носит название кодирование текстовой информации.

При кодировании текста чаще всего используется двоичная система счисления. Так как операции по данной системе чрезвычайно просты, то позволяют легко реализовать в технических устройствах известный принцип «все» или «ничего» («да» или «нет»). Большинство ЭВМ работают по такому принципу.

Так, например, этот принцип используется при осуществлении перевода текста, написанного на естественном языке, в текст, который может быть прочитан на ЭВМ. Каждая буква, цифра и знак в двоичной системе счисления изображаются в виде набора двоичных цифр 0 и 1. Эти комбинации двоичных единиц с помощью специальных аппаратов переносятся на перфоленту или перфокарты, которые представляют собой особую субстанцию плана выражения.

Так как преобразование текста из его естественной формы в машинную осуществляется при участии человека-оператора, то скорость и безошибочность подготовки исходного лингвистического материала к вводу в ЭВМ зависят от быстродействия и индивидуальных способностей оператора. Кодирование больших массивов приводит, безусловно, к механическим ошибкам, которые выявляются благодаря двух и трехкратным сверкам контрольной ленты с оригиналом. Весь этот процесс (перфорирование) подобен работе машинистки и тем самым нарушает принцип экономии, связанный в автоматизацией лингвистических исследований. По свидетельству составителей частотного словаря сочетаемости современного английского языка, над которым работают около 30 кафедр вузов страны, для получения статистически достоверных данных о грамматической и лексической сочетаемости 35 наиболее частых слов необходимо обработать текст длиной не менее 5 млн. словоупотреблений. Оказывается, перфорация лингвистически обработанного текста такой длины потребует около 5 млн. 45-разрядных перфокарт и 100 лет работы одного перфораторщика [1, 24].

В своей практике работы на ЭВМ «Минск–22» мы в основном были заняты кодированием текстовой информации на пятипозиционну бумажную перфоленту (п/л.) Следует отметить, что п/л имеет свою независимую от языкового материала физическую структуру, которая накладывает ограничений на кодируемый лингвистический материал подобно тому, как графическое построение алфавита или акустико-физиологическая структура

накладывает свои ограничения при кодировании лингвистической информации. Ограничения проявляются в виде искажений текста, связанных с нарушением физических свойств п/л.

Закодированная п/л является материальным носителем графической субстанции языковой формы. Субстанция осуществляется на п/л в виде комбинаций пробивок, согласно коду М-2. Пробивки в ЭВМ преобразуются в электромагнитные сигналы. Тем самым языковая форма представляется уже в виде новой субстанции. Она, эта последняя, в свою очередь декодируется на выводе ЭВМ, чтобы придать ей на бумаге вид исходной графической субстанции языковой формы.

Таким образом, предмашинный процесс работы требует от исследователя большого сосредоточения и широкий диапазон знаний, заключающий в себе подготовку как по инженерной, так и по традиционной лингвистике. Тщательное редактирование обрабатываемого текста, а затем и его качественное кодирование решают успех статистического анализа языкового материала на ЭВМ.

Особенности кодирования современных письменных текстов тюркских языков. Аппараты, кодирующие текстовую информацию, вводные и выводные устройства современных ЭВМ основаны на латинской и русской графике. Как известно, многие тюркоязычные народы СССР, принявшие русскую графику, включили в свои алфавиты, с целью отражения фонематического состава родного языка, дополнительные буквы с диакритическими знаками. В это число входит и казахский язык. Так как основная цель данной работы имеет в виду применение ЭВМ для автоматического анализа тюркских языков, то рассмотрим специфику их кодирования в предмашинный период.

Подавляющее большинство тюркских языков мира наличествует в Советском Союзе, они составляют 19 письменных языков [2, 28-34]. Каждая из этих письменностей имеет специфические фонемы, обозначающиеся русскими буквами с добавлением диакритических значков. Для примера можем указать на следующие начертания: заднеязычный глухой смычный «***қ***» – в казахском, каракалпакском, узбекском и уйгурском языках; заднеязычный носовой сонант «***ң***» – в башкирском, казахском, тувинском, туркменском, уйгурском, хакасском языках; переднерядный губной среднего подъема «***ö***» и «***ө***»-почти во всех тюркских языках и т.д. Исключение составляют крымско-татарский, кумыкский и ногайский языки, где специфические фонемы выражаются уже двумя буквами, как – ***гъ, дж, нъ, къ***, ***иг, оъ, уъ***. Двухбуквенные обозначения специфических фонем, наряду с выражением одной буквой, снабженной диакритический знаком, встречаются в балкарском, карачаевском, якутском языках. С учетом всех этих отличий письменностей тюркских языков нами пересмотрены и их алфавиты. Для удобства кодировки текстов и для универсальности машинных программ по обработке лингвистических материалов этих языков мы в своей исследовательской практике применяли единые обозначения идентичных фонем различных тюркских языков. Они передаются той или иной буквой

без диакритического значка, но с добавлением в правом верхнем углу апострофа «‘» или в сочетании со знаком «+». Исключение составляют буква «***i***» в казахском, хакасском языках, «***ə***» в азербайджанском, башкирском, казахском, татарском, туркменском, уйгурском языках, «***j***», в авербайджанском, алтайском языках, которые имеют однобуквенный вид, т.е. «***ə***» – «***э***», «*i*» – «1», «***j***» – «1». Так, встречающаяся в словах не тюркского происхождения буква «э» в нашей кодировке будет иметь вид «***э***», т.к. подобных слов в тюркских текстах сравнительно немного. Имеющиеся в алфавите двухбуквенные обозначения специфических фонем оставлены без изменения. И, наконец, вместо отсутствующего в коде машины твердого знака «***ъ***» взят «***ь***». Наша транскрипция современных тюркских письменностей преследует узкую цель – ввод текстовой информации в машину. Принятые нами буквозаместители для обозначения специфических фонем позволяют беспрепятственно пользоваться кодировочным аппаратом, основанном на русской графине (см. статью в сборнике «Статистика казахского текста». Алма-Ата, 1973, стр. 257).

Обозначение одной фонемы двумя символами имеет и свои недостатки, заключающиеся:

а) в увеличении количества символов, т.е. в увеличения объем текста;

б) в необходимости особого блока распознавания этих буквозаместителей в машине;

в) в непривычности чтения (на первых порах) лингвистического материала на выходе ЭВМ.

Но в целом подобные затруднения в обработке текстов тюркских языков в системе «**человек-машина-человек**» не дают повода для скептического отношения к делу, а наоборот, по сравнению о обработкой текста без применения ЭВМ являются весьма незначительными.

Для проверки на практике единой транскрипции тюркских языков мы составили экспериментальные частотные и алфавитно-частотные словари по казахскому, узбекскому, туркменскому, азербайджанскому, киргизскому и каракалпакскому языкам, объёмом от 2 до 10 тысяч словоупотреблений по каждому [3].

На основе данной транскрипции нами получены также частотные списки букв и буквосочетаний (2-х, 3-х) казахского текста, как с учетом их позиций в словах, так и без них.

Кроме вышеназванной транскрипции специфических букв, нами разработан принцип транскрибирования, который позволяет ввод в ЭВМ фонетического текста. Исследование фонетической структуры языка способствует определению состава фонем, частотности индивидуальных фонем, а также рядов позиционных вариантов фонем и т.д.

Благодаря предложенной транскрипции, были отмечены фонетические особенности вариантов казахских фонем, таких как твердость и мягкость, лабиализованность и нелабиализованность согласных и т.д. Подобно буквенной транскрипции, в основу фонетического транскрибирования положены двухсимвольные обозначения фонем. Теоретической основой для

нашего исследования послужили транскрипции перехода от графемного уровня и фонетическому, разработанные К.Б.Бектаевым и С.К.Кенесбаевым [4; 5]. На базе данной машиной транскрипции мы составили также фонетический алфавит узбекского, каракалпкского, ногайского языков, отражающий фонологические закономерности орфоэпии.

Как известно, рунические тексты древнетюркских письменных памятников (орхонские, енисейские) стали достоянием научной общественности, благодаря усилием таких ученых, как В.Томсен, В.В.Радлов, С.Е.Малов и др. [6; 7]. Членами группы статолингвистики разработана система индексов, позволяющих уяснить характер дистрибуции графических знаков и установить закономерности их позиционной транскрипции. На сегодняшний день тексты орхонских памятников закодированы (переведены на перфоленты) согласно нашей транскрипции и получены частотные списки по всем намеченным аспектам.

ЛИТЕРАТУРА

1. Ф.А.Каганов, Г.П.Мельников, В.И.Перебейнос, С.И.Хмельницкий. Автоматический ввод лингвистической информации в машину. Сб. «Автоматическая переработка текста методами прикладной лингвистики». Кишинев,1971.
2. К.М.Мусаев. Алфавиты языков народов СССР. М.,1965.
3. А.Х.Джубанов. Универсальная программа получения на ЭВМ «Минск-22» алфавитно-частотного списка сочетаемости графем. Сб. «Материалы второй научной конференции молодых ученых АН Каз ССР». Алма-Ата, 1970.
4. Ж.А.Аралбаев, К.Б.Бектаев, А.А.Исенгельдина и др. Статистическое исследование фонетической структуры казахского языка. СИИТЯ, Алма-Ата,1969.
5. І. Кенесбаев, Ғ. Мусабаев. Қазіргі қазақ тілі. Алматы,1962.
6, С.Е.Малов. Памятники древнетюркокой письменности. М. 1951.
7. Г.Айдаров. Язык орхонских памятников древнетюркской письменности VIII века. Алма-Ата, 1971.

ЗАДАЧА ПОЛУЧЕНИЯ НА ЭВМ ЧАСТОТНЫХ СПИСКОВ ЛИНГВИСТИЧЕСКИХ ЕДИНИЦ

АН Каз.ССР Институт языкознания.
Труды группы «Статистико-лингвистическое исследование и автоматизация»:
«Статистика казахского текста».
Алма-Ата, «Наука». 1973. С. 263-287.

Общий алгоритм задачи построения частотного списка. Выявление частот встречаемости и составление списков лингвистических единиц,

упорядочение их по интересующему нас признаку (по частоте, по алфавиту, по длине и т.д.) играет первостепенную роль в статистическом анализе изучаемого текста. Это дает возможность решения основисй задачи по отбору наиболее важных формальных единиц конкретного подъязыка и определению их информационных характеристик, помогает в построении формальной части морфологии, лексики и синтаксиса базового языка (сокращенного варианта естественного языка) [1]. Частотные списки можно получать как вручную, так и машинным способом.

Рассмотрим в общих чертах последовательность операций, выполняемых человеком при получении частотного списка без применения ЭВМ. К ручной выборке человек приступает только после уточнения задачи, уяснив себе необходимую для анализа лингвистическую единицу. В процессе чтения он отбирает интересующие его единицы и фиксирует их путем выписывания на карточки. В результате накапливается известное количество карточек, содержащих необходимые данные. Далее следует операция перебора и сортировки карточек на основе тех или иных признаков лингвистических единиц и определения их количества по группам. После чего составляется список анализируемых единиц: сначала записываются единицы, представляющие максимальное чиоло всех карточек (если идет сортировка по частоте), указывается их число (частота), а далее запись идет в порядке убывания частоты. Полученный таким образом список лингвистических единиц называется частотным списком (частотным словарем).

В принципе составление частотных списков «ручным способом» не является сложным. Но многократное повторение одних и тех же механических (не творческих) операций отнимает много времени и утомляет исследователя, что приводит к всевозможным ошибкам и неточностям. Поэтому такую часть умственной работы желательно поручить «электронному помощнику» – ЭВМ.

При машинном способе составления частотного списка также предварительно уточняется, что надо считать. Причем, последовательность перечисленных выше ручных операций повторяется почти в том же порядке, независимо от характера частотных списков. Для чего необходимо алгоритмизировать процесс автоматической обработки лингвистических данных. Заметим, что методика решения конкретной задачи разрабатывается, исходя из принципов решения уже известных смежных задач.

Дадим некоторые разъяснения понятия алгоритма. Алгоритмами называются точные предписания по выполнению в известном порядке определенной системы операций, необходимых для решения всех типовых задач. Алгоритмы делятся в свою очередь на принципиальный алгоритм и подалгоритмы [2, 7 и сл.стр.].

Принципиальным алгоритмом называется последовательность алгоритмов, записанная в виде схемы, приводящая к решению определенной

сложной задачи. Подалгоритм – это алгоритм решения отдельной простой задачи, входящей в состав сложной.

Алгоритмы и подалгоритмы записывают различными способами: специальными символами, блоками и т.д. В нашей работе мы использовали запись алгоритмов в виде блоков. Блок – это часть алгоритма или подалгоритма, которая может быть представлена одной или несколькими командами ЭВМ[1].

Общий алгоритм построения частотного вписка можно представить в виде следующей укрупненной блок-схемы, приведенной на рис. 1.

Кратко остановимся на функциях каждого из блоков. Ограниченность «памяти» современных ЭВМ не позволяет вводить в них за один прием достаточно большие текстовые материалы. Поэтому необходимо разбивать тексты на порции. Так, в нашей группе «Статистка речи» практикуется деление текста не порции, по 1000 словоупотреблений в каждой. Это соответствует объёму памяти ЭВМ «Минск-22») которая одновременно должна запомнить программу задачи составления частотного словаря, исходную информации (1000 словоупотреблении) и промежуточные результаты обработки – первичные частотные списки. В связи с этим в блоке 1 осуществляется ввод: сначала – первой тысячи словоупотреблений с перфоленты, затем, после ее обработки, второй тысячи словоупотреблений текста и т.д. до полной обработки необходимого объёма. Эта часть программы – общая для всех задач по составлению частотных списков. Отличительной частью программы являются те процессы, которые осуществляются во втором блоке.

Переведенный на перфоленту текст представляет собою множество комбинаций отверстий. ЭВМ, читая текст с перфоленты, воопринимает его как комплекс чисел в двоичной системе счислений (изображения чисел с помощью только «0» и «1»). Если для краткости чтения перенести каждое двоичное число в восьмеричную систему счисления, то можно убедиться в том, что отдельным буквам, цифре и другим символам соответствует двузначное восьмеричное число (А-70, Б-68 и т.д.). Таким образом, в памяти машины изучаемый текст запоминается как цепочки чисел. Выделение же нужной лингвистической единицы или распознавание лингвистического образа осуществляется благодаря определенному признаку, свойственному только данной единице. Например, выделение любой буквы текста или определенной заранее заданной буквы, цифровых или буквенных индексов, слогов, морфем, словоупотреблений, определенных словосочетаний, фраз, знаков препинания и, наконец, целого предложения можно осуществить только по заранее известным признакам.

---

[1] Отдельная команда несет в себе информацию об операции, подлежащей выполнению, и о величинах, над которыми должна выполниться эта операция, а также сведения о том, куда должен быть помещен результат выполнения операции. Система операции каждой конкретной машины называется системой команд данной машины. В систему команд входят операции арифметические, логические, операции управления и некоторые специальные операций.

Алгоритмизация каждой задачи по составлению частотных списков определяется содержанием второго блока, в котором происходит выделение необходимой лингвистической единицы. Оперативная память машины способна запомнить частотные списки ограниченной длины.

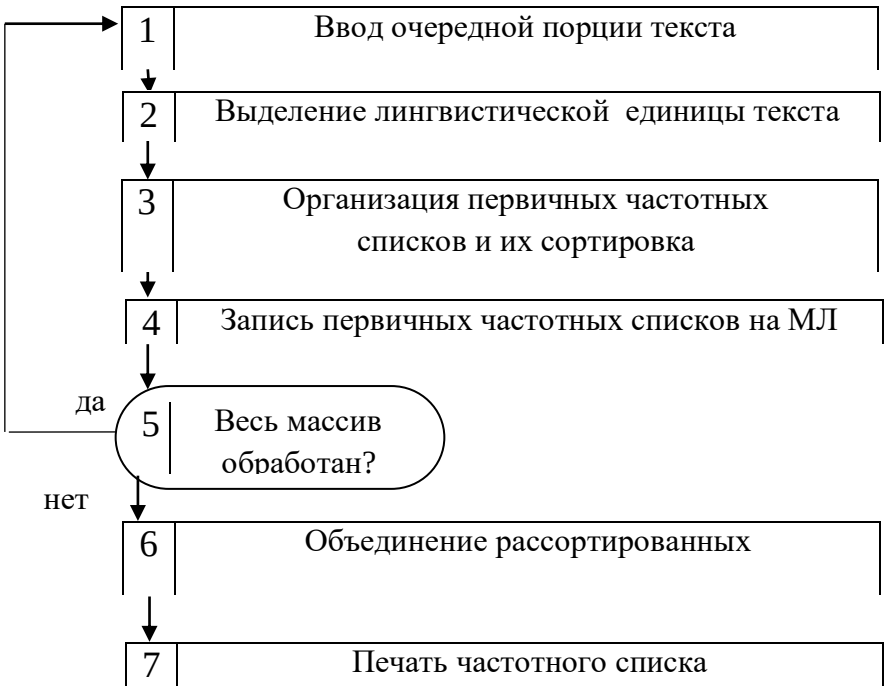


Рис.1. Общий алгоритм задачи построения частотного списка лингвистических единиц

Поэтому при обработке текста большого объема сначала составляются первичные частотные списки. Это осуществляется в ***блоке 3***, а процесс записи каждого первичного списка во внешнюю память (МЛ) – в ***блоке 4. Блок 5*** контролирует завершение обработки необходимого объёма текста. Если имеются еще не обработанные порций, то управление передается вновь на ***блок 1***, в обратном случае – на ***блок 6***, где осуществляется объединение первичных частотных списков в единое целое при сортировке по какому-либо признаку. Полученный словарь сохраняется во внешней памяти машины. Распечатка словаря с его статистической обработкой осуществляется ***блоком 7***.

Для того чтобы ЭВМ могла выполнять действия, указанные в алгоритме, необходимо их свести к тому набору операции, который может осуществлять вычислительная машина, и записать в такой форме, в какой машина может воспринимать и «понимать» операций, т.е. необходимо перейти от принципиального алгоритме, к алгоритму, записанному языком операции машины, закодировать последний алгоритм – записать его в виде последовательности команд. Полученная таким образом последовательность

команд называется программой, а сам процесс перехода от исходного алгоритма к программе называется программированием.

Именно на этом этапе решения лингвистических задач особенно заметно разделение труда между лингвистами и математиками-програмистами. В атом отражается не только вполне естественное разделение квалификации между ними, но также деление между информацией о тексте, написанном на естественном языке, которая имеет силу безотносительно к машине, и техникой передачи этой информации к конкретной машине. Как показывает опыт работы группы «Статистика речи», наибольший эффект в общении человека и машины может быть достигнут, когда в одном человеке сочетается лингвистическая подготовка с некоторой квалификацией в области программирования [3, 316].

После того как алгоритмы лингвистической задачи запрограммированы, наступает этап реализации всех программ на ЭВМ. Для чего в начале проводят отладку программы, т.е. выявление с помощью ЭВМ логических недоработок в процессе составления программы или ошибки в передаче переходов алгоритма в конкретных адресах машины, или, наконец, ошибок в процессе кодирования программы или исходной информации. Процесс отладки – это не что иное, как диалог человека и машины. Аварийные остановки машины – это сигнал машины о непонимании текста, а исправления – реакция человека на запрос машины, т.е, как бы перефразировка исходного задения, направленная на установление взаимопонимания человека и машины.

Решение задачи по отлаженной программе есть автоматическая реализация построенной модели, результаты которой показывают, насколька построенная модель точно отражает те или иные свойства оригинала.

На сегодняшний день алгоритмизация задачи построения частотных списков словоформ разработана в достаточной мере. Проблемы создания частотных словарей явились специальной темой обсуждения ряда всесоюзных конференций [4; 5; 6; 7; 8], освещались на страницах печати и, следовательно, не требуют подробного рассмотрения в данной работе. Однако нужно отметить, что имеющиеся программы по получению частотных списков словоформ нуждаются а частичном пересмотре в связи с особенностями тюркских языков, которые имеют свои специфические буквы, а также свой порядок расположения букв по алфавиту.

Учитывая эти особенности, мы в процессе работы составили машинный алфавит для целого ряда тюркских языков (казахский, узбекский, туркменский, азербайджанский, киргизский, каракалпакский и др. см. табл.1, 3, 4 в кн. «Статистика казахского текста» стр. 288-291).

Основываясь на общем алгоритме задачи построения частотного списка лингвистических единиц, приводам объединенный алгоритм задачи получения трех видов частотных словарей словоформ (см. рис. 2). В рассматриваемом алгоритме построения частотных словарей словоформ лингвистической единицей подсчета является словоупотребление, т.е. любая цепочка букв (и др.символов) между двумя пробелами. Алгоритм составлен

из расчета, что максимальная длина слова равняется 24 символам. На этой основе в блоке 2 нашего алгоритма осуществляется образование из очередного слова машинного документа, который состоит из пяти ячеек памяти машины (ЭВМ «Минск-22»). В первых четырех ячейках размещается слово, в пятой – частота его употребления. Необходимость подведения информации к подобной стандартной структуре объясняется простотой дальнейшей обработки (особенно при сортировке информации).

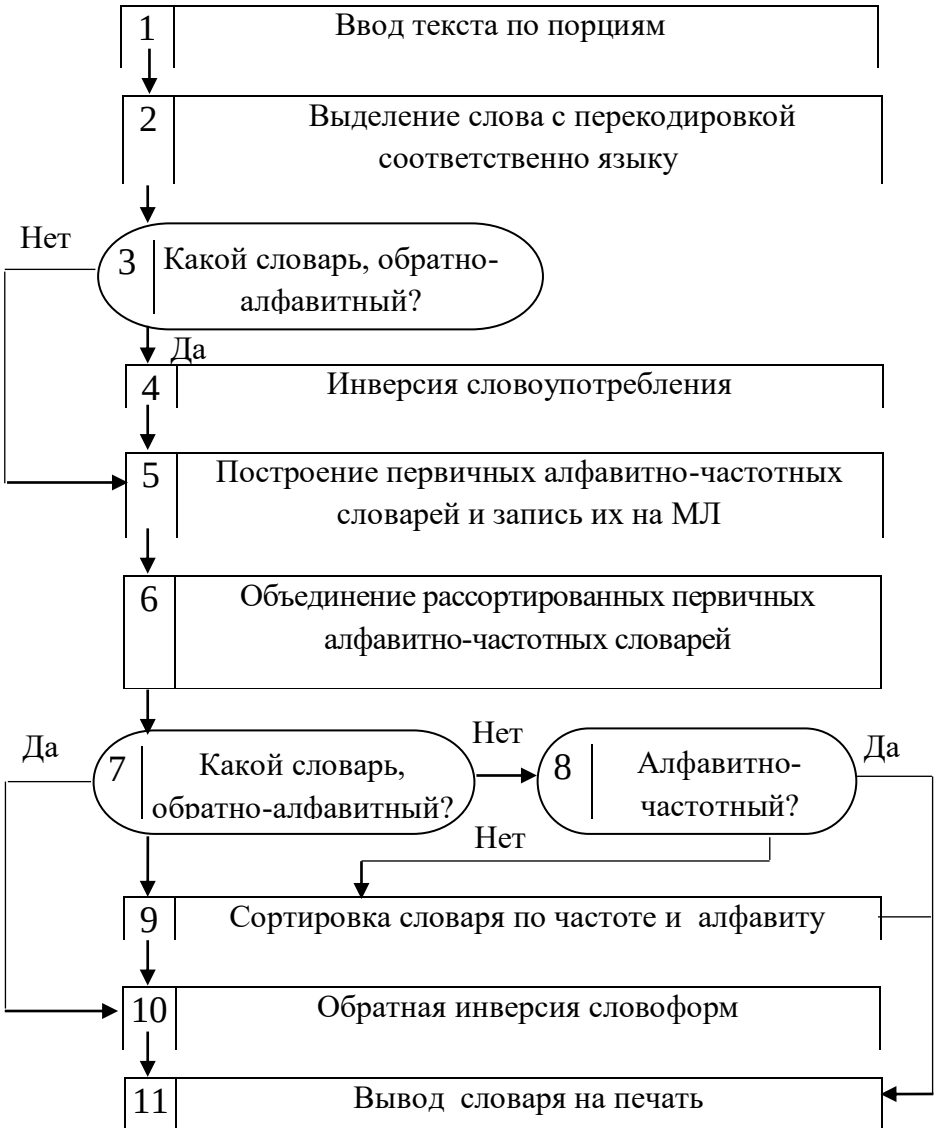


Рис.2. Общий алгоритм задачи получения частотных словарей словоформ

Полученные нами таким образом частотные словари словоформ представляют собой: I) алфавитно-частотный словарь, где словоформы упорядочены в строгом соответствии с алфавитом и с указанием частоты встречаемости каждой словоформы в тексте, 2) частотно-алфавитный словарь, т.е. список словоформ, упорядоченных по убыванию частоты каждой словоформы в тексте, а словоформы, имеющие одинаковые частоты, упорядочены еще и по алфавиту, 3) обратно-алфавитный частотный словарь, где словоформы упорядочены по алфавиту, начиная с конечных букв.

Во всех этих словарях каждая зафиксированная словоформа снабжается показателями абсолютной частоты, накопленной абсолютной частоты, относительной частоты и другими статистическими характеристиками [9].

Принципиальней алгоритм построения частотного списка индексов и их сочетаний. В этом параграфе дается принципиальный алгоритм задачи построения частотного списка, когда «лингвистической единицей» являются сами индексы. Для решения такой задачи не важно, какой именно индекс, т.е. определяет ли он лексические, морфологические и синтаксические признаки слова. Важен факт, что словоупотребления в тексте после себя имеют какие-то пометы – индексы. Результатом решения задачи может быть:

а) частотный список индексов;

б) Частотный список сочетаний из 2-х, 3-х, до 6-ти индексов. Второй случай характерен только для сплошной индексаций текста.

Данный алгоритм учитывает частный случай, когда после каждого слова имеются только один индекс. Случай, когда количество индексов, определяющих признак данного слова, будет больше одного, нами специально не расматривался, но небольшая модернизация алгоритма решает и такую задачу.

По приведенному на ***рисунке 3*** принципиальному алгоритму вкратце изложим назначение основных его блоков.

Блок 1 служит для настройки программы на каждую вновь решаемую задачу, куда входят подготовка (очистка) нужного количества рабочих ячеек, формирование необходимых констант и т.д. Назначения ***блоков 2–5*** схожы с назначениями ***блоков 1, 2*** и ***5***, рассмотренных в алгоритме предыдущего параграфа, т.е. ввод текстовой информаций и контроль за полной обработкой очередной порции текста или всего массива. Чтобы отличить индексы от конечных букв словоупотреблений, не имеющих индекса, мы решили проставлять перед буквой (или цифрой), овначающей индекс, знак тире «-» . Благодаря чему в ***блоке 6*** осуществляется выделение последних двух символов из слова и сравнение их с перечнем принятых индексов, имеющих знак тире впереди себя. При их совпадении выделенный из слова индекс отделяется от знака тире и приписывается к ранее выделенным (подобным же образом) индексам до заполнения всех разрядов ячейки (***блок 7***). Если данная ячейка не имеет свободных разрядов, то выделенный индекс записывается в разряды следующей по адресу ячейки. Когда расчитанное заранее количество ячеек памяти полностью заполнится, происходит запись содержимого таких ячеек на соответствущие зоны магнитной ленты (Мл).

Несовпадение выделенных 2-х символов с перечнем индексов (вместе с тире) означает отсутствие индекса в данном слове.

Таким образом, роль рассмотренной части алгоритма заключается в преобразовании первоначальной исходной информации в новый массив, состоящий из индексов, посредством которых соблюдается последовательность слов. При сплошном индексировании встречающиеся в

тексте знаки препинания (запятые, точки и т.д.) так же, как индексы, включаются в массив. Учет этих знаков может быть полезным при позиционном измении индексов с точки зрении синтаксического порядка слов.

Во второй части алгоритма организуются первичные частотные списики индексов или сочетаний из 2-х, 3-х, ..., 6-ти рядом расположенных индексов в зависимости от нашего выбора. Настройку работы программы на соответствующее число индексов в сочетании осуществляет ***блок 9***. Блоки 10–13 выполняют такую же функцию, как и рассмотренные блоки 2–5 в первом части алгоритма. Только в ***блоке 10*** вместо ввода текста (Блок 2) происходит чтение индексов с МЛ. ***Блок 11*** осуществляет вызов с содержимого каждой отдельной ячейки где хранятся индексы, и т.д.

Рис. 3 (начало). Принципальный алгоритм построения частотного списка индексов и их сочетаний

Неоходимое количество сочетаний индексов организуется в ***блоке 14***. Если требуется получить частотный список одних только индексов, то вся задача этого блока сводится к выделению очередного символа из цепочек индексов.

Каждый раз, блогадаря программе сортировки по алфавиту (***блок 15***), выделенный индекс располагается в общем списке: а) увеличивая на единицу частоту аналогичного индекса; б) в случае отсутствия такового, находя соответствующее место в списке согласно алфавиту.

Посроение частототных списков 2-х, 3-х, 4-х, 5-и, 6-и индексных сочетаний отличается от вышеописанного процесса тем, что после выделения очередного символа нужно приписывать его к ранее выделенным до тех пор, пока не получится необходимое количество индексов в сочетании. Если, например, в одной из ячеек памяти оказались такие индексы, как *ЕЗЗГСО*, то организованные из этого количестве 2-х индексные сочетания будут: ЕЗ, ЗЗ, ЗГ, ГС, СО, 3-х индексные: ЕЗЗ, ЗЗГ, ЗГС, ГСО; 4-х индексные: ЕЗЗГ, ЗЗГС, ЗГСО; 5-и индексные ЕЗЗГС, ЗЗГСО; 6-и индексные: ЕЗЗГСО. Число различных вариантов при любом случае сочетаний можно подсчитать по следующей простой формуле: $\mathbf{N = K^n}$, где **N** – число всевозможных вариантов; **К** – число различных индексов, т.е. количество присваемых признаков; **n** – количество индексов, необходимое в сочетании.

Рис. 3 (конец)**.** Принципальный алгоритм построения частотного списка индексов и их сочетаний

Например, подсчитаем число **N**, когда индексы будут означать принадлежность слов к одной из части речи. При учете вплоть до знаков препинания (запятая и точка) число различных индексов **К**=11. Далее, **n**-может быть (для нашей программы) равно 1,2,3,4,5,6. Тогда:

$\mathbf{N} = 11^1 = 11$ $\mathbf{N} = 11^2 = 121$ $\mathbf{N} = 11^5 = 161051$

$\mathbf{N} = 11^3 = 1331$ $\mathbf{N} = 11^4 = 14641$ $\mathbf{N} = 11^6 = 1771561$

Если теперь взять К>11, то при таких же значениях «n» будем иметь соответственно большие значения **N**. Таким образом, оба случая, как увеличение числа различных индексов, так и увеличение числа индексов в сочетаний приводят к увеличению числа всевозможных вариантов сочетаний, т.е. к «удлинению» частотного списка.

Учет указанных обстоятельств необходим для распределения памяти ЭВМ. Если при построении частотного списка по нашему алгоритму окажется, что емкость машины недостаточна (ОЗУ для получаемого списка будет исчерпана), то ***блок 15*** перепишет полученный первичный частотный список из ОЗУ в определенные зоны МЛ. Третий этап работы нашего алгоритма – объеднение (если в этом есть необходимость) рассортированных первичных частотных списков, записанных на МЛ, что входит в функции ***блока 17***.

Наконец, благодаря ***блоку 18*** объединенный алфавитный список, полученный на МЛ, печатается на АЦПУ.

Принципальный алгоритм построения частотного списка букв и буквосочетаний. Имеющиеся программы по построению на ЭВМ частотных списков букв и буквосочетаний не были рассчитаны для текстов тюркских языков, которые имеют ряд специфических букв. Основной отличительной чертой предлагаемой нами программы является блок распознавания букв соответствующего тюркского алфавита. Как известно, для текстов большинства индоевропейских языков свойственно соответствие каждого отдельного символа слова определенной букве. Это облегчает процесс выделения букв. Кроме того, в этих программах весь обрабатываемый текст рассматривается как одно «длинное слово», что не совсем соответствует нашим заиндексированным текстам. Таким образом, еще одна отличительная черта нашей программы – это наличие блока распознавания слов, не имеющих индекса, от слов с индексом и различение слова от индекса (если таковой имеется).

Алгоритмитизация этой задачи, конечно, не представляет особой сложности, но необходимость составления ее при исследовании тюркских языков на графемно-фонемном уровне бесспорна.

Перечислим возможности программы, составленной для ЭВМ Минск-22» по алгоритму приведенному на **рис.4**.

1. Исходной информацией могут служить тексты, написанные на любом из тюрских явыков, тексты, основанные на русской и латинской графике, а также транскрибированный фонетический текст.

2. Построение следующих алфавитно-частотных списков графем:

а) алфавитно-частотный список сплошных[2] сочетаний графем с учетом пробела между словами. Количество букв в сочетании для текстов не тюркских языках равно от 2-х до 4-х; для текстов на индоевропейских языках от 2-х до 6-ти[3];

б) алфавитно-частотный список сочетаний графем в начальной и в конечной позиции слова (количество букв в сочетания такое же, как и в пункте «а»);

в) алфавитно-частотный описок сочетаний графем, равноотстоящих от границ (концов) слова или, как мы условно называем, сочетаний букв в средней позиции (количество букв в сочетании такое же, как в пункте «а»).

Данный случай требует некоторых пояснений. Для того чтобы организовать сочетание, соотоящее из нечеткого количества букв, необходимо, чтобы и само слово состояло из нечетного числа букв и, наоборот, для образования сочетания из четного количества букв, слово должно состоять из четного числа букв. Например, из слова «***жұмысшы***», состоящего из 7 букв, возможны такие сочетания: однобуквенное – «***ы***», трехбуквенное – «***мыс***», пятибуквенное – «***ұмысш***» и семибуквенное «***жұмысшы***». Из слова «***мектеп***» (6 букв) можно организовать такие четные сочетания в средней позиции: двухбуквенное – «***кт***», четырехбуквенное - «***екте***» и шестибуквеное – само слово «***мектеп***» и т.д. Во всех частотных описках указывается абсолютная, абсолютно-накопленная и относигельная частоты.

3. Побочным продуктом программы является: частотный список – «количество букв в словах»; частотный описок – «количество слов в предложении».

Кратко поясним назначение каждого из блоков алгоритма данной программы (***рис.4***).

В ***блоке 1***, в зависимости от выбора пускового адреса команд, соответственно числу сочетающихся букв, происходит подготовка необходимых констант и очистка рабочих ячеек. ***Блок 2*** осуществляет поочередный ввод порций исходной информации в память машины. Дальнейшая передача управления зависит от положения ключей, указанных в инструкции. Дело в том, что используемый для кодировки текстов международный код ***М-2*** не позволяет в дальнейшем осуществлять

---

[2] Под сплошным, сочетанием графем надо понимать,что после первой организации необходимого количества сочетании букв из текста во все последующие будет входить предыдущие сочетания без первой буквы. Например, трехбуквенные сплошные сочетания из фразы «***біздің үй***» будут иметь следующий вид: первое сочетание – ***біз***, второе – ***ізд***, третье – ***зді*** и последующие – ***дің, дүй*** и т.д.

[3] Такое разграничение языков в зависимости от количества букв в сочетании возникло вследствие того, что запись любого количества сочетаний производится в одной только ячейке. Для ЭВМ типа «Минск» в одной ячейке можно записать не более шести букв.

сортировку по алфавиту. В связи с этим необходимо перекодировать каждую букву алфавита данного языка, поставив в соответствие с возрастающим или убывающим рядами чисел в порядке их расположения в алфавите.

Новый промежуточный код (**ПК**) для разных тюркских языков, а жакже для русского и латинского алфавитов приведен в табл. 2,3 4 (см. сб. «Статистика казахского текста», 1973, стр. 289-291).

Перекодировка из иода М-2 в новый ПК, соответствующий алфавиту исследуемого языка, осуществляется ***блоками 3, 4, 5***. Словоупотребления и знаки препинания текста при перфорации отделяются друг от друга знаком пробела, т.е. единицы текста отграничиваются от непрерывных цепочек двухзначных (восьмеричная с/с) чисел.

Таким образом, пробел и точка дают возможность выделить нужную единицу (слово, предложение), а также по мере необходимости вести счет слов, предложений и фиксировать количество слов в каждом предложении.

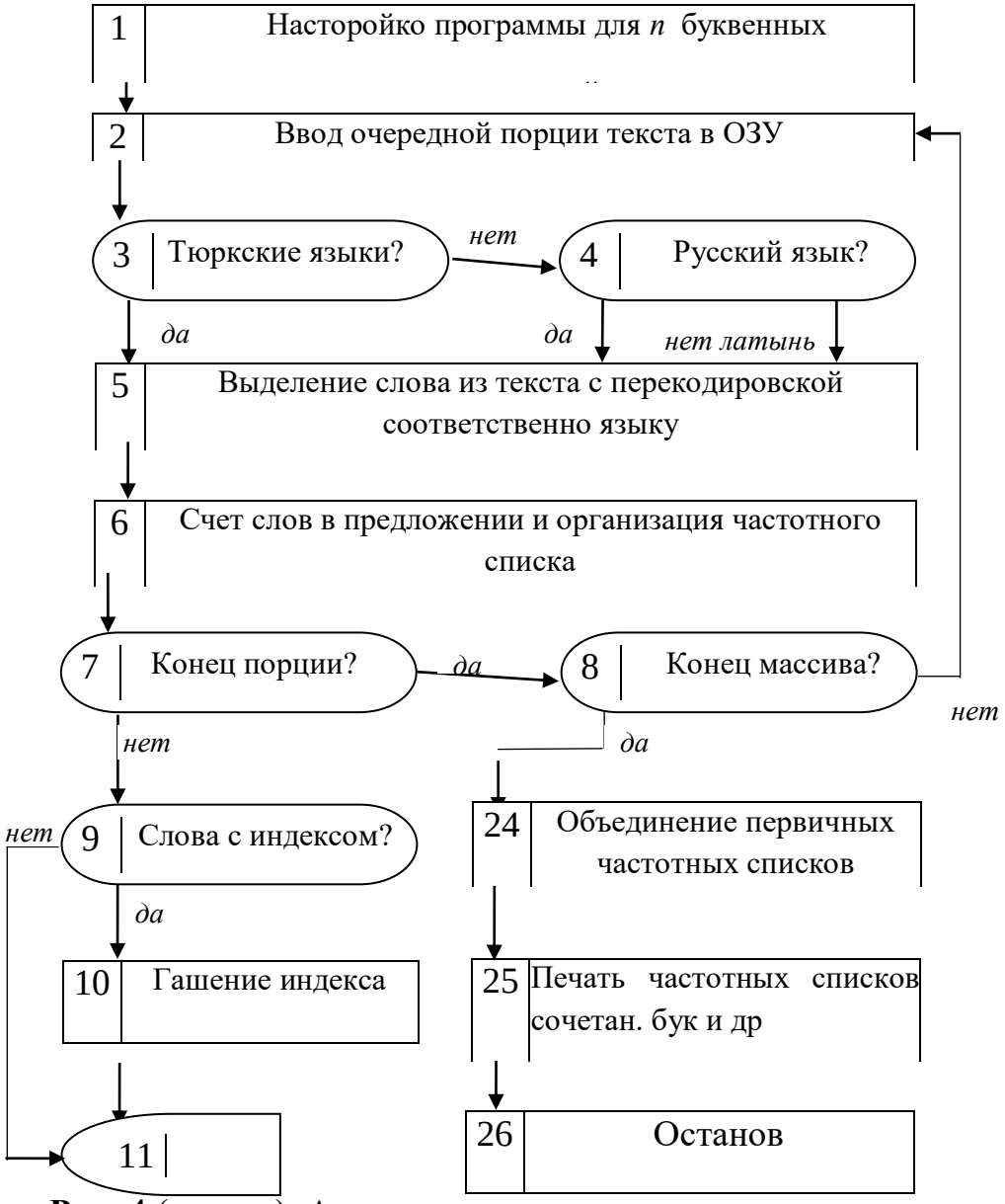


Рис. 4 (начало). Алгоритм задачи построения частотного списка букв и буквосочетаний

Блок 6 строит частотный список по количеству слов в предложениях (т.е. количество предлоянений, состоящих из одного или двух, трех и т.д. словоупотреблений). Как уже отмечалось, словоупотребления в тексте могут быть как с индексом, так и без него. Контроль за этим осуществляет ***блок 9***. При наличии индексов управление передается ***блоку 10***, где происходит гашение индекса. От задачи к задаче количество индексов может меняться, и тогда с пульта управления необходимо занести в определенную ячейку (указанную в инструкции) нужную константу.

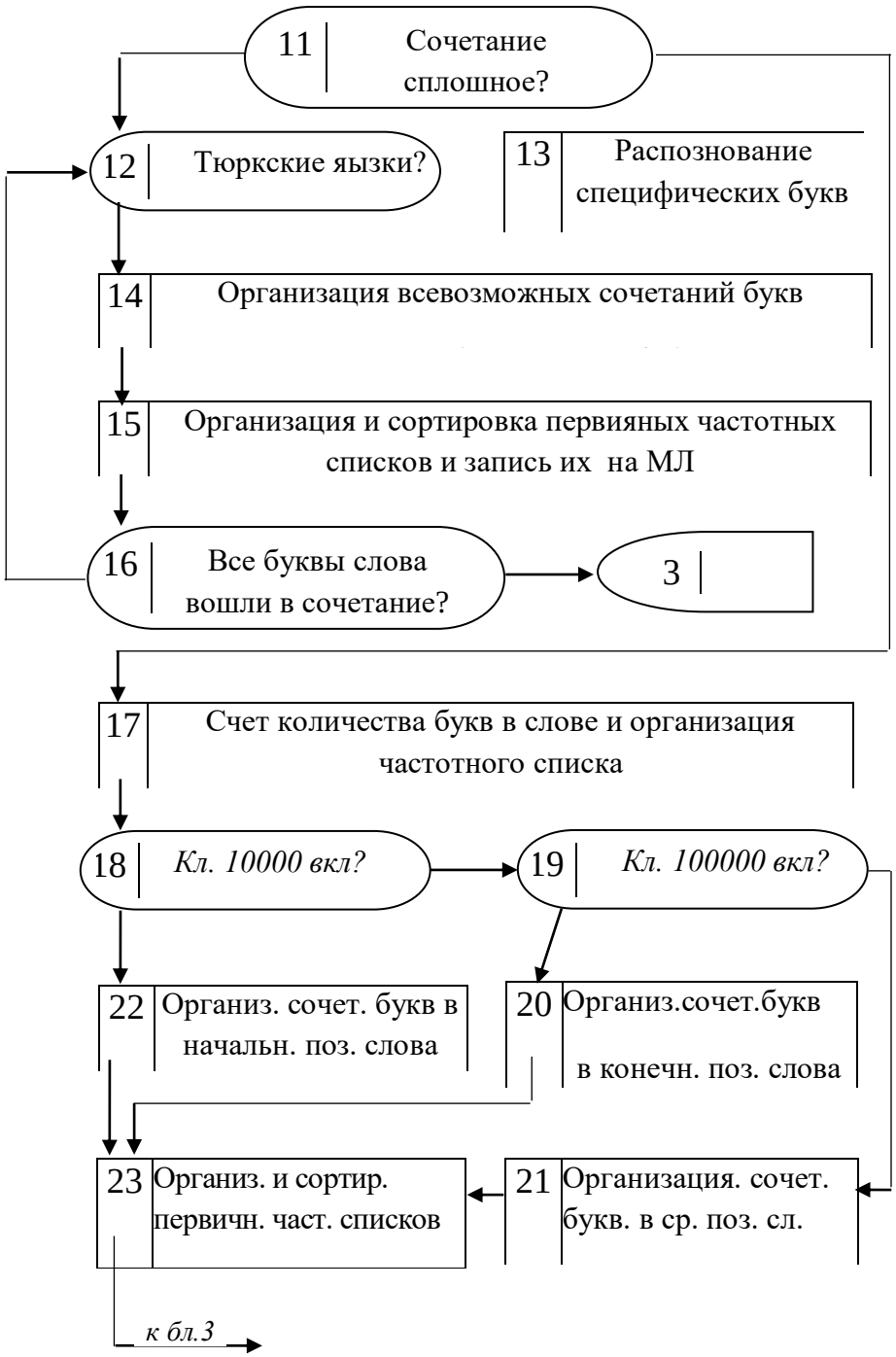


Рис. 4 (конец). Алгоритм задачи построения частотного списка букв и буквосочетаний

После гашений соответствующего количества индексов так же, как и при отсутствии индекса у слова, управление передается ***блоку 11***, который, в зависимости от положения соответствующего ключа может передать управление или блоку построения частотных списков сочетаний букв по позициям.

Каков же путь получения сплошного сочетания? Когда исходной информацией являются тексты на тюркских языках, включается блок распознавания специфических букв. Далее и в случае, когда обрабатываются тексты на индоевропейских языках, управление передается ***блоку 14***, где осуществляется организация сочетаний с соответствующим количеством букв. Построение первичных частотных списков букв и буквосочетаний с одновременной сортировкой по алфавиту входит в назначение ***блоков 15*** и ***16***.

Получение частотных списков букв и их сочетаний по позициям в слове осуществляется с помощью блоков, начиная от ***17*** по ***23*** (и некоторых других, ранее рассмотренных). По условию задачи количество букв в слове должно быть больше необходимого количества сочетающихся букв или же равно им. Поэтому в ***блоке 17*** вначале осуществляется подсчет количества букв в слове. Причем для текстов на тюркских языках подключается блок распознавания специфических букв. Одновременно, так же, как и при построений частотного списка «количество слов в предложении», составляется частотный список «количество букв в слове». Такой частотный список указывает на количество однобуквенных, двухбуквенных, трехбуквенных и т.д. до 24-х буквенных слов в данном тексте.

Блоки 18 и ***19*** в зависимости от положения указанных ключей передают управление или ***блоку 22***, или ***блоку 21***, или ***блоку 20***, в которых организовываются сочетания букв в начальной, средней и конечной позиций слов соответственно.

Подалгоритм задачи построения частотных списков сочетаний букв по позициям в слове несколько подробнее показан на **рис.5**, который не требует особых описаний.

После фиксации блоком 8 (рис.4) законченности обработки исходной информации начинается процесс объединения первичных частотных списков, записанных на МЛ, если в этом есть необходимости. Одновременно происходит упорядочение по алфавиту. Последним этапом является вывод алфавитно-частотного списка букв или буквоссчетаний со статистической обработкой. Фрагменты частотных списков, полученных на основе данной программы показаны в ***табл. 7, 8, 9, 10*** (см. сб. «Статистика казахского текста», 1973, стр. 294-297).

Принципальный алгоритм построения частотного списка интервалов между тождественными буквами. Появление той или иной буквы в последовательной цепи какого-либо текста является случайным событием. Стало быть, определение вероятности последующего появлении букв имеет важное значение для целей прикладной лингвистики, а именно – в автоматическом распознавании речевых образов. Нами поставлена задача

алгоритмизировать получение частотного спика интервалов между тождественными буквами. В понятие интервала входит количество букв, пробелов, знаков препинания, заключенных между идентичными буквенными знаками. Так, частотный список интервалов между двумя «***а***» может содержать следующие сведения:

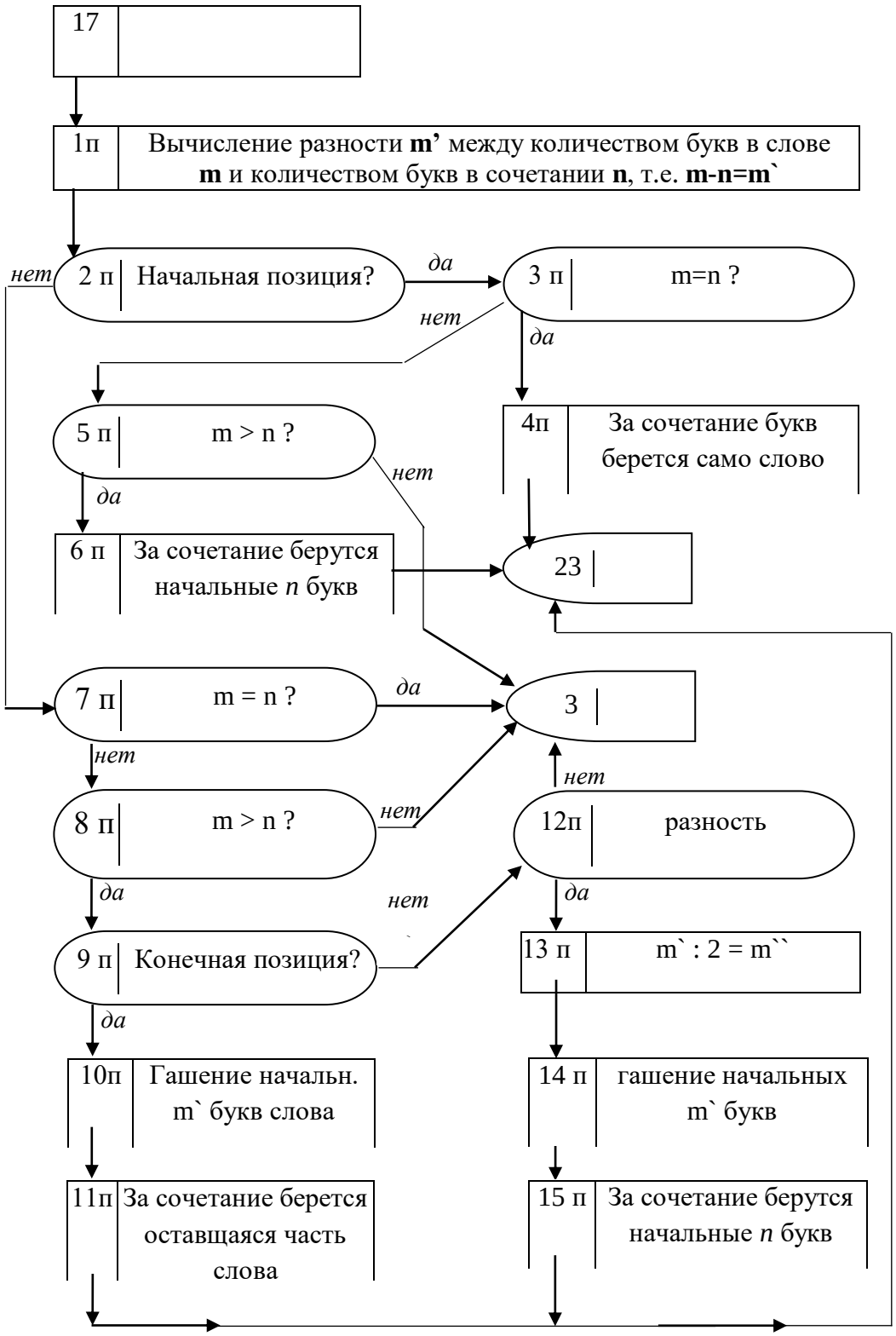


Рис. 5. Подалгоритм задачи построения частотных списков сочетаний букв по позициям в слове

а) частота встречаемости интервала, равного «нулю», т.е. когда «***а***» непосредственно следует за «***а***»;

б) частота встречаемости «***а***» через один интервал;

в) частота встречаемости «***а***» через два, три и т.д. интервала. Алгоритмизация такой задачи не является сложной. В алгоритме ее блоки, с 1 по 6, повторяют функции уже рассмотренных аналогичных блоков предыдущих алгоритмов (см.**рис.6**). Отличительной чертой данного алгоритма является единица счета, представленная количеством интервалов.

Рис. 6. Алгоритм задачи построения частотного списка расстояний в количествах букв (интервалов) между одноименными буквами

Определение подобной переменной величины входит и функцию блоков 7 и 8. Назначение остальных блоков указано в ***блок-схеме 6*** (рис. 6) данного алгоритма. Фрагмент частотного списка интервалов между тождественными буквами приведен в ***табл.11*** (см. сб. «Статистика казахского текста», 1973, стр. 298).

Литература

1. Р.Г.Пиотровский. Базовые языки и отраслевой бинарный перевод. «Язык и общество», Л.,1968.
2. Б.А.Трахтенберг. Алгоритмы и машинное решение задач. М., 1960.
3. А.В.Зубов. Переработка текста естественного языка в системе «Человек-машина». Сб. «Статистика речи и автоматический анализ текста». Л., 1971.
4. Межвузовская конференция по вопросам частотных словарей и автоматизации лингвистических работ. Тезисы докладов и сообщений. ЛГУ. Ленинград, 1966.
5. Частотные словари и автоматическая переработка лингвистических текстов. Минск,1968.
6. Статистическое и информационное изучение тюркских текстов. Тезисы докладов и сообщений. Алма-Ата, 1969.
7. Струнтурно-математические методы моделирования языка. Тезисы докладов всесоюзной конференции. Киев, 1970.
8. Автоматическая переработка текста методами прикладной лингвистики. Материалы всесоюзной конференции 6-8 октября 1971. Кишинев, 1971.
9. К.Б.Бектаев, А.Х.Джубанов, А.В.Зубов. Автоматическое построение частотных словарей (прямого и обратного). «Вестник АН Каз ССР» 1970, Каз ССР». 1970, №3.

ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ТЮРСКИХ ТЕКСТОВ

АН Каз.ССР Институт языкознания.
Труды группы «Статистико-лингвистическое исследование и автоматизация»:
«Статистика казахского текста».
Алма-Ата, «Наука». 1973. С. 299-328.
(соавтор – К. Б. Бектаев).

Статистический метод широко применяется в индоевропейском языкознании с конца прошлого столетия. В области тюркского языкознания он не был столь распространен, если не считать отдельных исследований, основанных на количественном подходе к языковым явлениям, суть которых сводится в основном к установлению процентного содержания лингвистических данных на том или ином уровне языка.

Мы считаем, что реальное основание для количественного изучения тюркских языков и их свойств заложены в самой их языковой структуре. Выражаясь словами Р.Г.Пиотровского, «Бурные ассимиляции, редукции и пр., что входят в понятие «фонетических законов» в рамках индоевропейского языка, чужды тюркским языкам, а это дает возможность

точно и недвусмысленно фиксировать те или иные моменты языка при математическом описании».

Задачей нашего исследования является теоретическая разработка и реализация универсальных схем автоматического построения вероятностно-статистических моделей распределения лингвистических единиц в тюркоязычных текстах.

Создание вероятностно-статистических моделей необходимо, главным образом, для автоматического поиска информации, оптимизации, преподавания языков, создания вероятностного базового языка, решения теоретической задачи о вероятностной природе «нормы», для изучения вопросов стилистики, а также при решении многих других задач.

В нашем исследовании предпринимается попытка автоматического построения вероятностно-статистической модели распределения классов слов на материале казахского текста.

Прежде чем перейти к решению этой задачи, обратимся к описанию тех приемов математической статистики, с помощью которых осуществляется анализ данных в соотнесении их с моделями теории вероятностей.

Статистический эксперимент и вероятностная модель. Проверка лингвостатистических гипотез. Статистический анализ результатов количественного лингвистического эксперимента является важнейшим звеном в построении математических моделей языка. Благодаря ему мы получаем возможность выявить ряд свойств языка, скрытых от непосредственного наблюдения.

Построенные дедуктивным путем теоретико-вероятностные модели характеризуются определенными количественными параметрами. Например, для **нормального распределения** таковыми будут математическое ожидание M(X) и средне-квадратическое отклонение σ.

Предполагается, что с помощью вероятностных моделей удается описать количественное поведение некоторых лингвистических единиц в тексте. Если же принять эту характеристику как показатель действительности, то возможно будет перейти от модели к самому объекту. Для этого необходимо определить количественные параметры исследуемых текстовых единиц, как среднее арифметическое $\overline{X}$ или стандарт S, которые можно сравнить с аналогичными параметрами вероятностных моделей. Причем указанные параметры интересуют нас не в связи с конкретным текстом ограниченной длины, а как характеристика генеральных совокупностей текстов (язык в целом, стили, подъязыки, язык писателя), т.к. эти совокупности из-за своего объема не могут быть полностью подвергнуты статистическому анализу. Следовательно, объектом исследования служат отдельные частные выборки, взятые из генеральных совокупностей. Таким образом, генеральная совокупность выступает в роли натурного объекта, а частные выборки – как модели. С помощью параметра $\partial_в$ частной выборки, применяя математический аппарат, можно оценить параметр $\partial_г$ генеральной совокупности. Результаты этого позволяют выяснить правомерность

распространения выводов, полученных при исследовании частной выборки, на всю генеральную совокупность. Выявленные индуктивным путем параметры выборочной совокупности сопоставляются с параметрами дедуктивных вероятностных моделей для определения оптимальной модели, интересующей нашу генеральную лингвистическую совокупность.

Таким образом, статистико-математическое исследование частной выборки и ее параметров относительно генеральной совокупности является средством, дающим возможность осуществить переход от вероятностной модели к лингвистическому натурному объекту. Оно осуществляется на основе определенных критериев (правил) по проверке статистических гипотез.

Прикладная лингвистка, как и всякая другая наука, развивается благодаря возникновению и проверке всевозможных гипотез. По своим формулировкам и методам их проверки эти гипотезы весьма разнообразны. Среди них особую роль играют статистические (лингвостатистические) гипотезы, которые формулируются либо относительно вида, либо относительно параметров распределения случайной величины.

Так можно выдвинуть следующую гипотезу; распределение существительных в тексте – **нормальное**. Чтобы оценить степень совпадения нашего эмпирического распределения с теоретическим, т.е. **нормальным распределением**, необходимо обследовать выборочные значения средних частот существительных по различным сериям. Выдвинутая гипотеза о виде или о параметрах распределения случайной лингвистической величины, как известно, может быть сформулирована как нулевая (или основная) гипотеза $\mathbf{H_0}$ или как альтернативная гипотеза – $\mathbf{H_1}$, противосоставленная $\mathbf{H_0}$.

В упомянутом выше случае утверждение о том, что распределение существительных в тексте является **нормальным**, представляет собой нулевую гипотезу $\mathbf{H_0}$.Что касается выдвигаемого против нее антитезиса, который оспаривает **нормальность распределения** существительных, то он являет собой альтернативную гипотезу – $\mathbf{H_1}$. Разумеется, что альтернативных гипотез может быть и несколько. Однако нами рассматриваетя лишь тот случай, когда нулевой гипотезе – $\mathbf{H_o}$ противостоит одна единственная альтернативная гипотеза – $\mathbf{H_1}$, хотя она может состоять в свою очередь из нескольких гипотез подобного рода. Итак, наша альтернатива отрицает нулевую гипотезу, т.е. $\mathbf{H_1}=\hat{\mathbf{H}}_0$. Заметим, что проверка нулевой гипотезы производится с помощью применения определенного критерия. Сами критерии различаются как собственно статистические и порядковые критерии. Первые из них проверяют нулевую гипотезу, а для чего обращаются к распределению значений случайной лингвистической величины, как, например, частота и частотность, а вторые – рассматривают упорядоченность этих величин.

Если гипотеза H_0 касается значений тех или иных параметров распределения, то статистические критерии, проверяющие ее, именуются параметрическими. Критерии, применяемые для оценки гипотез о характере распределения, также относятся к разряду параметрических. Так для того

чтобы проверить гипотезу о **нормальней** или **пуассоновском распределении** глаголов, мы должны сравнить параметры эмпирического и теоретического распределения.

Наоборот, целый ряд критериев не требует обращения к значениям параметров. Их мы называем непараметрическими. Они просты, но менее эффективны.

Лингвостатистические гипотезы, как правило, проверяются на случайной выборке. Однако такая выборка не отражает с идеальной точностью распределение генеральной лингвистической совокупноси. Следовательно, она может направить процесс проверки лингвостатистичоской гипотезы и по ложному пути. Вследствие подобного заблуждения возможны ошибочные выводы двоякого рода: во-первых, нулевая гипотеза H_0 отвергается, если даже она верна. Таким образом, принимается ложная гипотеза H_1. Во-вторых, принимается ложная нулевая H_0, отвергающая тем самым правильную альтернативную гипотезу H_1.

Может возникнуть вопрос, не произвольно ли определяется одна из двух возможных гипотез как нулевая, т.е. основная, а другая – как альтернативная? А также: не является ли чисто условным соглашением то, что одно из неверных решений мы относим к ошибке первого рода а другое – второго?

Ответим на эти вопросы. Выбор гипотез зависит от обстоятельств, ибо для исследователя небезразлично, какую из них принимать за основную. Этим самым определяется и степень значимости ошибок первого и второго рода. Отметим, что значимость ошибок, особенности в лингвистике, устанавливается часто путем субъективных умозаключений. К примеру, т.н. алтайская гипотеза [1, стр. 23-41] о генетическом родстве тюркских, монгольских, тунгусо-манчжурских языков сформулирована таким образом, что проверить ее можно с помощью известного статистического критерия. Нулевая гипотеза здесь допускает, что все алтайские языки восходят к одному генетическому истоку, т.е. языку-основе. Отвергая ее, алтаист, даже самый убежденный, совершил бы опасную ошибку, которую мы назвали выше ошибкой первого рода. Здесь менее опасно ошибка второго рода, хотя сама гипотеза, принятая в противовес нулевой, будет неверна. Следовательно, выбор критерия проверки гипотезы в то же время является и выбором критической области.

Дадим понятие критической области, основываясь на теоретико-множественных представлениях из аксиоматических построений теории вероятностей А.Н.Колмогорова [2].

Пусть U обозначает множество (пространство) значений случайной величины X, а W есть выборочное подмножество этих значений, которые мы наблюдали в опыте. Оно, т.е. W, состоит из возможных выборочных точек e_1, e_2,..., e_m . Каждая из них представляет собой определенную совокупность наблюдаемых значений случайной величины x_1, x_2,..., x_n . К тому же точки e_1, e_2,...,e_m рассматриваются как некоторые выборочные числовые значения некой статистической характеристики E, относительно которой выдвигаются

гипотезы H_0 и H_1. Нулевая гипотеза H_0 проверяется с помощью некоторого статистического критерия «С». Эту гипотезу можно отвергнуть в том случае, если наблюдаемая характеристика, т.е. выборочная точка e_i , попадает внутрь определенной области подмножества, в так называемую критическую область Wc. Когда наблюдаемая характеристика попадет внутрь области, которая носит название области приемлемости решении гипотеза H_0 принимается.

В заключение следует отметить, что при принятии того или иного решения важно в большей степени не допускать ошибки первого рода, нежели второго. В целях уменьшения степени ошибок и производится испытание гипотез. Необходимо помнить о том, что выбор статистических критериев, а следовательно и критической области, мы должны осуществить таким образом, чтобы выборочные наблюдения e_i как можно реже попадали в критическую область Wc . Тогда нулевая гипотеза H_0 окажется справедливой. Именно такую цель преследует произвольная фиксация малого числа α, которое указывает на уровень существенности и выдвижение требования, согласно которому при проверке истинной гипотезы H_0 ошибка первого рода не должна превалировать над этим уровнем.

Исходя из практических соображений, т.е. из ожидаемых последствий ошибки первого рода, определяем свой выбор уровня существенности. Говоря по-иному, в зависимости от характера последствий необходимо уменьшить и уровень существенности. Обычно в лингвистических исследованиях применяется $\alpha = 0{,}05$.

Данный уровень существенности, например, считается вполне приемлемым для статистической проверки алтайской гипотезы.

К вопросу о законах распределения классов слов. В настоящее время лингвостатистика от количественного описания лингвистических явлений, именуемых первым этапом ее развития, переходит на следующий, более сложный этап. Этот этап должен включать построение теории, позволяющей предсказывать количественно те или иные закономерности, получать «**уравнение лингвистики**» [3, 44].

Таким образом, проблема установления вида распределений единиц текста является одной из наиболее важных задач модельного построения лингвистических явлений. Этому вопросу посвящен ряд исследований, в частности, С.Г.Чебанова, Р.М.Фрумкиной, В.М.Калинина, Р.Г.Пиотровского, Г.Хердана, Фукса и др. [4-19].

В предлагаемой статье осуществляется проверка гипотез о соответствии статистических распределении предлагаемым теоретическим распределениям (конкретно: нормальный закон, логарифмически-нормальный, распределение Пуассона и распределение Шарлье в двух типах – А и В).

Статистические распределения существительных, глаголов, прилагательных и наречий строились нами на материале выборочно отобранного художественного текста (романа М.Ауэзова «Абай жолы») объемом 10 тысяч словоупотреблений и 50 тысяч словоупотреблений

публицистического (газетного) текста. В данной статье впервые предпринимается попытка автоматического построения вероятностно-статистической модели распределения классов слов на материале казахских письменных текстов. Поэтому сначало опишем методику сформирования статистических (эмпирических) распределений и вероятностные модели лингвистических случайных величин. Далее, осуществим проверку гипотез о соответствии эмпирического распределения с теоретической с тем, чтобы найти подходящую для них вероятностную модель.

Построение эмпирического распределения частот. При изучении характера изменения каких-то лингвистических единиц (в нашем случае классов слов) немаловажное значение имеет знание минимального объема текста, начиная с которого закономерность изменения не нарушалась бы. Нами сделана попытка определения минимального объема однородного текста и числа серии, начиная с которого изменение (варьирование) частоты встречаемости частей речи подчинялось бы каким-либо из упомянутых законов распределения. Для чего общая выборочная совокупность (или просто выборка) строилась следующим образом.

Каждая словоформа выборочно-отобранных текстов была отмечена определенной буквой, обозначающей классы слов.

Далее, по программе на ЭВМ производилось деление текста на серии (микровыборок) объемом 25, 50, 100 и т.д. словоупотреблений. Возможные количества таких микровыборок из наших текстов, а также полученные общие объемы приведены в ***таблицах 1*** и ***2*** . Для краткости записи примем такие обозначения:

Таблица 1

	Значения «***n***» по роману		
N \ **K**	25	50	100
2500	100	–	–
5000	200	100	–
10000	400	200	100

Таблица 2

	Значения «***n***» в публицистическом тексте				
K \ N	25	50	100	200	500
2500	100	–	–	–	–
5000	200	100	–	–	–
10000	–	200	100	–	–
12500	500	–	–	–	–
20000	–	–	200	100	–
25000	1000	500	–	–	–
40000	–	–	–	200	–
50000	2000	1000	500	250	100

К – объем микровыборки (серии); n – количество микровыборок;

N – общий объем текста в каждом из случаев;

B_N^n=***К*** – общий вид записи отдельного случая, которая читается: из текста объема ***N*** случайно отобраны ***n*** выборок (микровыборок), каждая из которых состоит из ***К*** лингвистических единиц (знак равенства «=» является условным).

На ЭВМ «Минск-22» по составленной нами программе получен частотный описок индексов (индексы определяют классы слов) для каждой серии объемом 25, 50, 100, 200 и 500 словоупотреблений.

Для отдельного варианта B_N^n=***К*** подсчитывалась частота встречаемости анализируемой единицы внутри «микровыборок» *К*. После чего определялось количество серий, в которых анализируемая единица встречалась с одной и той же частотой. Далее строились дискретные вариационные ряды распределения, состоящие из перечня встречавшихся различных частот и числа микровыборок, соответствующих этим частотам, в виде следующей таблицы:

x_i	x_1	x_2	x_3….	x_s
n_i	n_1	n_2	n_3….	n_s

где x_i (i = 1, 2, ..., s) – различные частоты исследуемого признака, встретившиеся в микровыборке ***К***; n_i ($i=1, 2,...,S$) – количество микровыборок с частотой $\boldsymbol{x_i}$. Причем,

$$\sum_{i=1}^{S} n_i = n, \qquad (1).$$

Построенный таким образом дискретный вариационный ряд частот мы используем для проверки соответствия распределений классов слов в текстах теоретическим законам распределения Пуассона и Шарлье типа В. При проверке соответствия распределений классов слов в текстах с другими теоретическими законами (нормальным законом, логарифмически-нормальным и законом Шарлье типа А) мы строили непрерывные вариационные ряды путем преобразования дискретных вариационных рядов. Для чего производилось деление эмпирического ряда на укрупненные классы. В зависимости от размаха варьирования и от условно принятых количеств классов выбирается длина интервала **ΔX** с помощью простой формулы [20, 126]:

$$\Delta \mathbf{X}=(\mathbf{X_{max}}-\mathbf{X_{min}}):\mathbf{r}\ , \qquad (2)$$

где $\mathbf{X_{max}}$, $\mathbf{X_{min}}$ – максимальная и минимальная частоты встречаемости анализируемой единицы. Разность их и определяет размах варьирования случайной величины. **r** – условно взятое число классов, зависящее от объема внутрисерийной выборки. Значение r определяется приближенно по следующей **таблице 3** [21, 20]:

Таблица 3

К	25	50	100	200	500
r	5	7	9	10	12

В качестве примера построения дискретных и непрерывных вариациационных рядов рассмотрим распределение существительного в художественном тексте по роману «Абай жолы». Рассмотрим случай, когда объем текста N=10000 словоупотреблений, объем микровыборок или серий К=100 словоупотреблений и количество таких серий n=100, т.е. случай $B_N^n = \hat{A}_{10000}^{100}$=***100***. После подсчета существительных в каждой из 100 микровыборок оказалось, что самая минимальная частота встречаемости равна 19, а максимальная – 43. Выяснив, в скольких микровыборках встречаются частота 19, 20, 21, 22, ..., 43, строим дискретный вариационный ряд, как показано в ***таблице 4***.

Таблица 4

x_i	19	20	21	22	23	24	25	26	27	28	29	30
n_i	1	0	2	3	3	3	2	8	2	8	10	14

31	32	33	34	35	36	37	38	39	40	41	42	43	Сумма
8	5	3	2	11	5	4	1	2	1	0	1	1	100

Для того чтобы преобразовать полученный дискретный вариационный рад в непрерывный вариационный ряд, поступаем следующим образом:

а) определяем размах варьирования:

$$\mathbf{x_{max} – x_{min}} = 43\text{-}19=24;$$

б) определяем число классов по ***табл. 3***, где для объема серии, равной 100, соответствует значение **r** =9;

в) вычисляем длину интервала: $\Delta \mathbf{X=(X_{max}–X_{min}):r}$=24:9=2,67.

Округлив полученное значение, будем иметь **ΔX**=3;

г) учитывая, что **ΔX**=3, наблюдаемые в микровыборке значения $\mathbf{X_i}$ группируем по 9 классам (интервалам), далее определяем значение каждого интервала (середину интервала): $\mathbf{X_i' = X_i + \Delta X:2;}$

д) для получения интервального ряда суммируем значения частот в пределах каждого интервала:

$$n_0' = n_0 + n_1 + n_2 = 1 + 0 + 2 = 3,$$

$$n_1' = n_3 + n_4 + n_5 = 9 \text{ и т.д.}$$

В результате приведенный к интервалу ряд принимает вид ***таблицы 5***.

Таблица 5

(x_i, $x_i+\Delta X$)	(19, 22)	(22, 25)	(25, 28)	(28, 31)	(31, 34)	(34, 37)	(37, γ_0)	(γ_0, γ_3)	(γ_3, γ_6)
x_i	20,5	23,5	26,5	29,5	32,5	35,5	38,5	41,5	44,5
n_i	3	9	12	32	16	16	7	2	1

Таким образом, построенные нами дискретные и непрерывные вариационные ряды будут являться эмпирическим распределением частот классов слов. В дальнейшем эти данные неоходимы при вычислении на ЭВМ теоретических частот и статистических параметров.

Вероятностные модели.

1. Закон Пуассона. Известно, что одним из наиболее важных распределений в лингвистическом исследовании является **Пуассоново распределение**. Оно имеет место в том случае, когда вероятность (***p***) отдельного наступления события очень мала, а число опытов очень велико. Предположим, что данное условие выполнимо и на нашем материале. Для проверки берем дискретные вариационные ряды, составленные по методу, изложенному выше. Когда случайная величина X подчинена распределению Пуассона с математическим ожиданием $\boldsymbol{\alpha}$, то теоретическая частота какого-либо значения «***x***» в «***n***» опытах определяется по формуле [20, 105]:

$$n^T = \frac{a^x \cdot l^{-a}}{x!} \cdot n, \qquad (3)$$

где через n^T обозначена теоретическая частота.

Мы в своей работе определяем средне-арифметическое значение частоты ($\bar{x}$), являющееся аналогом (моделью) математического ожидания:

$$a \approx \bar{x} = \frac{1}{n} \cdot \sum n_i \cdot x_i^1 \ , \qquad (4)$$

2. Распределение Шарлье типа В. Данное применяется в основном как приближенное к Пуассонову распределению [22, 308-314, 529-535].

Распределение Шарлье типа В в лингвистических исследованиях еще не применялось. Мы в своем эксперименте решили применить ее наряду с распределением Пуассона для проверки заданных гипотез. Теоретическая частота случайной величины, распределенной по закону Шарлье типа В, определяется по формуле:

$$n^T = n \cdot \psi(x) \cdot \left\{ 1 + \frac{\mu_2 - a}{a^2} \cdot \left[\frac{x(x-1)}{2} - ax + \frac{a^2}{2} \right] \right\} \ , \quad (5)$$

где ψ(x) – Пуассонова функция распределения, т.е.

$$\psi(x) = \frac{a^x \cdot l^{-a}}{x!} \ , \qquad (6)$$

μ_2 - центральный момент второго порядка, численно равный дисперсий, т.е.

$$\mu_2 = \sigma^2 = \frac{1}{n-1} \cdot \sum_i x_i \cdot (x_i - a)^2 \ , \qquad (7)$$

С помощью статистических характеристик α и δ^2 описываются важные стороны распределений, как, например, степень разбросанности значений переменной величины и положение кривой распределения на графике.

3. Нормальный закон распределения. Большое значение **нормального распределения** объясняется рядом обстоятельств. Нормальная кривая и ее свойства сравнительно несложны для применения. Нормальная кривая близка ко многим одновершинным распределениям, кроме того, для нормального распределения более подробно разработан целый ряд примеров в математико-статистической обработке материалов наблюдения. Наконец,

нормально распределенные совокупности имеют большое значение для определения объема выборок и оценок статистических характеристик.

Итак, если функция распределения φ(x) непрерывной случайной величины X, которая может принимать значения от - ∞ до + ∞, распределена по нормальному закону, то она выражается формулой

$$n^T = \varphi(x) \cdot n = \frac{n \cdot \Delta x}{\sigma} \cdot \frac{1}{\sqrt{2\pi}} \cdot l^{-\frac{(x-a)^2}{2\sigma^2}}, \qquad (8)$$

где ΔX – длина интервала, используемая при преобразовании дискретного ряда в непрерывный вариационный ряд, *α,* и *δ* – соответственно являются математическим ожиданием и дисперсией, которые определяются по формулам (4) и (7).

Для установления степени близости изучаемого распределения нормальному часто приходится вычислять асимметрию и эксцесс:

$$A = \frac{\mu_3}{\sigma^3} \qquad (9)$$

$$E = \frac{\mu_4}{\sigma^4} - 3 \qquad (10)$$

где σ – среднее квадратическое отклонение, μ_3, μ_4 – соответственно центральные моменты третьего и четвертого порядка, вычисляемые по формулам:

$$\mu_3 = \frac{1}{n} \cdot \sum_i n_i \cdot (x_i - a)^3 \qquad (11)$$

$$\mu_4 = \frac{1}{n} \cdot \sum_i n_i \cdot (x_i - a)^4 \qquad (12)$$

Эти характеристики для нормального распределения равны нулю. Поэтому, если для изучаемого распределения асимметрия и эксцесс мало отличны от нуля, то можно предположить, что такое распределение близко к нормальному. Наоборот большие значения асимметрии и эксцесса указывают на значительное отклонение изучаемого распределения от нормального.

4. Распределение, Шарлье типа А. Распределение Шарлье типа А является приближенным к нормальному закону распределения. Функция распределения его имеет следующий вид [22]:

$$n^T = \varphi(x) \cdot n = \frac{n \cdot \Delta\chi}{\sigma} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(x-a)^2}{2\sigma^2}} \cdot$$
$$\cdot \left\{ 1 + \frac{A}{\sigma}\left[\frac{(x-a)^3}{\sigma^3} - 3\left(\frac{x-a}{\sigma}\right)\right] + \frac{E}{24} \cdot \left[\frac{(x-a)^4}{\sigma^4} - \sigma\left(\frac{x-a}{\sigma}\right)^2 + 3\right]\right\} \qquad (13)$$

Выражение, стоящее перед фигурной скобкой, не что иное как нормальный закон распределения. Все параметры (13) совпадают с параметрами нормального закона распределения, их значения определяются по формулам (4), (7), (9), (10) .

5. Логарифмическо-нормальное распределение. В ряде случаев распределения, не являющиеся нормальными, можно В ряде случаев распределения, не являющиеся нормальными, можно преобразовывать в таковые, если заменить переменную x переменной $\sqrt{x}$ или ℓgx и. т. д [23].

В случае, когда переменная ***x*** в формуле нормального закона распределения заменена на $\boldsymbol{\ell gx}$, то его функция распределения называется логарифмически нормальным распределением и имеет вид:

$$n^T = n \cdot \varphi(x) = \frac{n \cdot \Delta x \lg e}{x \cdot \sigma} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(\lg x - \lg a)^2}{2\sigma^2}} \qquad (14)$$

где $\lg a = \frac{1}{n} \cdot \sum_i n_i \cdot \lg x_i$ – среднее арифметическое, (15)

а дисперсия – $\sigma^2 = \frac{1}{n-1} \cdot \sum_i n_i (\lg x_i - \lg a)^2$ (16)

Центральные моменты третьего и четвертого порядка равны .

$$\mu_3 = \frac{1}{n} \cdot \sum_i n_i \cdot (\lg x_i - \lg a)^3 \qquad (17)$$

$$\mu_4 = \frac{1}{n} \cdot \sum_i n_i \cdot (\lg x_i - \lg a)^4 \qquad (18)$$

Асимметрия, эксцесс, μ_3 и μ_4- для логнормального распределения (14) вычисляются по формулам (9), (10), (17), (18).

Оценка соответствия эмпирического распределения частот классов казахских слов теоретическому с применением χ^2 Пирсона. Каждый из выше перечисленных законов распределения является теоретическим законом или вероятностной моделью. Близость эмпирических распределений какому-либо теоретическому распределению отражает закономерные соотношения изучаемого признака. Следовательно, возникает необходимость оценки степени близости эмпирического распределения выбранному теоретическому закону распределения.

Такую оценку производим с помощью неоднократно применяемого в лингвистике критерия согласия χ^2 Пирсона [4, 5, 10, 11, 15-19, 24 и др.]. Данная процедура осуществляется следующим образом.

Выдвигается нулевая гипотеза **H₀**, заключающаяся в том, что эмпирическое распределение выборки соответствует теоретическому распределению генеральной совокупности.

Для того чтобы принять или отвергнуть гипотезу, конструируется некоторая статистическая характеристика:

$$W = \sum_{i=1}^{k} \frac{(n_i^э - n_i^T)^2}{n_i^T},$$

где $n_i^э$ и n_i^T – эмпирическая и теоретическая частоты распределения лингвистической единицы в тексте. Можно показать, что при N→∞ (объем выборки) статистическая характеристика W имеет распределение χ^2 со степенью свободы υ = t - ℓ - 1 (ℓ - число параметров линейных зависимостей, t - число групп эмпирического распределения).

Таким образом, соответствие эмпирического и теоретического распределений можно оценивать по расхождению эмпирических частот $n_i^э$ и математических ожиданий $n_i^T = N \cdot P_i$, где P_i - вероятность, соответствующая i-му классу, N – объем выборки. Тогда выражение:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i^э - N \cdot P_i\right)^2}{N \cdot P_i}, \qquad (19)$$

известно под названием критерия χ^2.

Если вычисленное значение χ^2=O, то считается, что эмпирические и теоретические частоты совпадают. В остальных случаях, когда χ^2 отличен от нуля, расхождение между этими частотами будет возрастать с удалением от нуля. Различают при этом существенное и несущественное расхождения. Установление этого факта решается после получения значения χ^2 и всех остальных параметров теоретических распределений.

Для оценки с помощью критерия χ^2 существенности и несущественности различия между эмпирическими и теоретическими частотами распределения, т.е. для проверки нулевой гипотезы H_0 необходимо:

а) определить число степеней свободы υ, которое равно числу групп эмпирического распределения ***K*** минус число параметров, найденных с помощью эмпирического распределения ***l*** без единицы, т.е. υ = ***K–l–1;***

б) выбрать уровень значимости *q%* для критерия. Например, значению вероятности 0,95 соответствует уровень значимости q=0,05 (или 5%). Это означает, что выход за пределы принятых границ возможен в порядке случайности с вероятностью q= 0,05 (5%) случаев могут быть ошибочными и т.д.;

в) найти табличное значение χ_q^2, являющееся функцией зависимости от числа степеней свободы и величины уровня значимости *q*: $\chi^2 = \chi^2$ (υ, *q*);

г) сравнить вычислительное значение χ^2 по данным выборки и найденному табличному значению χ_q^2 (пороговое значение). При этом мы будем иметь одно из двух неравенств – $\chi^2 \geq \chi_q^2$ или $\chi^2 < \chi_q^2$. В первом случае критерий попадает в критическую область и расхождение

выборочных данных с теоретическими существенно, тогда проверяемая гипотеза отвергается большие отклонения наблюдаемых частот от теоретических ведут к отклонению гипотезы. Во втором случае расхождение между теоретическими и эмпирическими данными несущественно гипотеза принимается. Оценку гипотезы H_0 можно осуществлять путем не только сравнения величин χ^2 и χ_q^2, но и знания величины вероятностей P($\chi^2 \geq \chi_q^2$).

Если P($\chi^2 \geq \chi_q^2$) окажется ниже некоторого заданного наперед уровня значимости, например ниже q=0,05, то это будет означать, что вероятность случайных отклонений очень мала. В этом случае придется признать нашу оценку отклонения эмпирического распределения от теоретического закона не случайной (случайные явления с очень малой вероятностью следует считать практически невозможными). Это будет указывать на неправдоподобность гипотезы о несущественности расхождений между эмпирическим и теоретическим распределениями. Если же вероятность P($\chi^2 \geq \chi_q^2$) достаточна велика (в нашем случае больше 0,05), то нулевую гипотезу о близости обоих видов распределений следует принять.

Значение такой вероятности указывает на степень отклонения или приемлемость гипотезы H_0 .

Известный специалист по математической статистике Г.Крамер, анализируя результаты опыта, считает совпадение хорошими при P($\chi^2 \geq \chi_q^2$)=0,17 очень хорошим при P($\chi^2 \geq \chi_q^2$)=0,91 и почти значимым отклонение от теоретического распределения при P ($\chi^2 \geq \chi_q^2$)=0,03. По В.Романовскому, вероятность заключенная между 0,1777 и 0,2700, хотя не мала, ко вместе с тем и не велика [26, 41].

Таким образом, в том случае, когда различие соответствия эмпирических и теоретических частот не существенны, представляется возможным отдать предпочтение той вероятности, которая больше, и считать ее в этом случае весьма надежной, а меньшее значение вероятности менее надежной.

Теперь рассмотрим последовательность проверки эмпирических распределений с предполагаемым теоретическим законом распределения по критерию χ^2.

I. Индексирование текста по частям речи.

2. Составление словаря по программе «Частотный список индексов» для отдельного случая B_N^n=***К*** .

3. Построение (вручную) вариационных рядов и подготовка их к вводу ЭВМ по разработанной нами схеме.

4. Вычисление на ЭВМ значений:

а) среднего арифметического ($\bar{x}$);

б) среднего квадратичного отклонения δ;

в) коэффициента асимметрии А и эксцесса Е;

г) теоретических частот $n_i^T = N \cdot P_i$, используя формулу, рассмотренные в §3 (т.е. по формулам (3), (5), (8), (13), (14).

д) χ^2 для каждой *i*-ой группы и их суммы по формуле (19).

5. Объеденение (если в этом есть необходимость) частот со значением меньше 5 и введение соответствующей поправки на сумму $\sum_{i=1}^{k} \chi^2$ (вручную);

6. Нахождение числа степеней свободы *v*=ℓ-k-1 (для нормального, логнормального распределения и Шарлье –В ℓ=2, для распределения Шарлье – А ℓ=4 и для распределения Пуассона ℓ=1).

7. Определение значения интеграла вероятностей P($\chi^2 \geq \chi_q^2$) из таблицы [27 или др.] по вычесленному χ^2 и с числом степеней свободы ***v***.

Результаты нашего эксперимента, т.е. распределения существительных, глаголов, прилагательных и наречий отражены в сб. «Статистика казахского текста», изд. в 1973 г. (см. таблицы 6-14 на стр. 318-326).

Анализ табличных данных показывает, что выбор нормировок (объем серии, количество серий и объем выборки) существенно влияет на характер распределения изучаемой лингвистической единицы.

Под нормировкой подразумевается следующее: один и тот же текст объема ***N*** можно разбить на серии ***K*** объемом по 100, по 1000, по 2000 и т.д. словоупотреблений и в соответствии в этим можно выбрать количество серий ***n***. Причем во всех случаях будет N=kn. В роли серии в лингвостатистике выступает не только определенное количество словоупотреблений, но и «длина» строк, количество страниц и т.д. Следовательно, когда идет речь о распределении той или иной лингвистической единицы, то задача ставится не вообще, а конкретно – с какого объема серии считать непротиворечивым согласие эмпирического распределения с выбранным теоретическим. Например, мы не можем утверждать, что словоформы распределяются по закону Пуассона, а можем сказать, что при определенных нормировках они описываются законом, Пуассона, а при других нормировках, допустим, нормальным законом [4, 24, 28] или, начиная с такого-то объема серии, описываются определенным вероятностным законом. Далее всякие сравнения параметров, а также содержательные лингвистические интерпретации можно провести уже на основе этих нормировок [4, 24]. Например, при большом параметре закона Пуассона можно пользоваться как нормальным так и пуассоновым приближением, а при малом – пуассоновым [28, 193]. Таким образом, нахождение наилучшего способа разбиения текста безусловно, имеет большое значение в модельном построении лингвостатистики.

Теперь проанализируем распределения каждой из частей отдельности.

Существительные и глаголы по сравнению с другими частями речи в казахских текстах являются самыми частыми. Так, например, в тексте романа «Абай жолы» 35,36% слов составляют существительные и 31,44% а –

глаголы. В публицистическом же тексте существительные – 45,19%, а глаголы – 21,86% [29] . Поэтому если считать, что результативность выборки удовлетворяет требованиям независимости, то вполне можно предположить, что эмпирические распределения указанных классов слов согласуются с нормальным законом. Проверку данной гипотезы мы осуществляем путем сравнения величин $P(\chi^2 \geq \chi_q^2)$ для различных комбинаций значений N, n, k .

В таблицах 6 и 7 отражены значения вероятностей $P(\chi^2 \geq \chi_q^2)$ соответствия эмпирического распределения существительных с их теоретическими распределениями, т.е. как нормальное, Шарлье типа А и логнормальное, при различных комбинациях N, n, K . Значения вероятностей таблицы 7 для наглядности представлены в виде таблиц 6 и 9, где они (вероятности) сгруппированы по постоянному значению n и переменным n, K (таблица 8) и по постоянному N и переменными, n, K (таблица 9) (см.сб. «Статистика казахского текста», изд. в 1973 г., табл. 6-14 на стр. 318-326).

Анализ значений вероятностей указанных таблиц показывает, что для всех рассматриваемых объемов серии К, т.е. для К=25, 50, 100, 200 словоупотреблений, объемом выборок N (2500, 5000, 10000, 12500, 20000, 40000, 50000 словоупотреблений) и различных количеств серий n=N:K, надежной моделью, описывающей распределение существительных, является нормальный закон распределения. Следующим по надежности при таком сравнении оказалось распределение Шарлье типа А. Для существительных газетного текста при объеме серии К≥100 вполне допустимо логнормальное распределение.

В таблицах 10 и 11 (см.сб. «Статистика казахского текста», изд. в 1973 г., табл. 6-14 на стр. 318-326) даны значения вероятностей $P(\chi^2 \geq \chi_q^2)$ степени соответствия эмпирического распределения глаголов таким законам, как нормальному, Шарлье типа А и логнормальному при различных N, n, K. В результате сопоставления возможных случаев эмпирического распределения глагола в художественном и публицистическом текстах убеждаемся в том, что нормальный закон, так же как и в предыдущем случае, может служить математической моделью их распределения. Необходимо отметить, что при К≤100 эмпирическое распределение глагола наилучшим образом соответствует закону Шарлье типа А. Нормировки: К≥100, n≥100 и N≥10000 для публицистического текста благоприятствует распределению глагола в соответствии с нормальным законом.

Опираясь на эти доводы, можно считать, что закон нормального распределения является моделью эмпирического распределения казахских глаголов (по крайней мере для указанных нормировок).

Анализ вероятностей совпадения распределений существительного и глагола с законами Пуассона и Шарлье типа В показал их явное несоответствие. В связи с этим в указанных таблицах значения их вероятностей отсутствуют.

Исследования на согласие эмпирических распределений казахских **прилагательных** художественного и публицистического текстов с рассматриваемыми теоретическими распределениями позволяют сделать следующие выводы.

Для значений объемов внутрисерийной выборки К≥25, для различных n≥100 и N≥2500 художественного текста и для К≥50, n≥100 и N≥5000 публицистического текста в подавляющем большинстве случаев распределение прилагательных согласуется с законами Шарлье типа А и В, а также с законом Пуассона. При такой ситуации предпочтение можно отдать Пуассонову распределению в силу его удобства и изученности. Однако, начиная с К≥200, n≥100 и N≥20000, нормальное распределение также хорошо согласуется с эмпирическим распределением прилагательных (см, таблицы 12 и 13 в сб. «Статистика казахского текста», изд. в 1973 г., на стр. 324-325).

Классы казахских слов, **наречий** (значения вероятностей приведены в таблице 14, там же стр. 326), чаще всего распределены в тексте по законам Пуассона и Шарлье типа В. Более удобным для применения, конечно, является Пуассоново распределение. Такое утверждение не означает, что при достаточно больших значениях объема внутрисерийной выборки, как, например при К≥500, предпочтение всегда отдается Пуаосонову распределению. Для таких нормировок, очевидно, весьма предпочтительным будет нормальный закон, т.к. в наших экспериментах хорошее согласие эмпирического распределения наречий с упомянутым теоретическим замечается при нормировке К≥100, n ≥100 и N ≥10000.

Таким образом, здесь предпринимается попытка найти наилучший способ разбивки текста, при которой эмпирические распределения казахских классов слов достаточно хорошо могут согласоваться с определенным теоретическим законом распределения. Результаты подобных исследований, на наш взгляд, имеют большое значение при изучении вопросов моделирования лингвистических фактов и явлений.

Литература

1. Дж. Клоусон. Лексико-статистическая оценка алтайской теории. ВЯ, 1969, №5.

2. А.Н.Колмогоров. Основные понятия теории вероятностей. Пер. с немецкого. М.-Л., 1936.

3. Р.Л.Добрушин. Математические методы в лингвистике. «Математическое просвещение». М., 1959, выпуск 6.

4. К.Б.Бектаев и К.Ф.Лукьяненков. О законах распределения единиц письменной речи. СРААТ, Л., 1971.

5. К.Б.Бектаев, А.В.Зубов, Е.Ф.Ковалевич, Л.Б.Машкина, О.А.Нехай. К исследованию законов распределения лингвистических единиц. СТІ, Минск, 1969.

6. К.Б.Бентаев, Р.Г.Пиотровский и др. О зоконах расспределения классов слов, словоформ и словосочетаний. Сб. Межвузовская конференция

по вопросам частотных словарей и автоматизации лингвистических работ. Л., 1966.

7. В.М.Калинин. Развитие схемы Пуассона и ее применение для описания статистических свойств речи. АКД., Л.,1964.

8. В.М.Калинин. Некоторые статистические законы математической лингвистики. Сб. «Проблемы кибернетики», М., 1964, №11.

9. В.М.Калинин. Математические аспекты восприятия иноязычного текста. Сб. «Статистика речи». Л., 1968.

10. К.Ф. Лукьяненков. Лексикостатистическое описание научно-технического текста с помощью электронно-вычислительной машины (подъязык судовых механизмов). КД (рукоп.). Минск, 1969.

11. Р.М.Фрумкина. О законах распределения слов и классов слов. Сб. «Структурно-типологические исследования». М., 1962.

12. С.Г.Чебанов. О подчинении речевых укладов «ивдоевропейской» группы закону Пуассона. Доклады АН СССР, новая сер., т.55, 1947, № 2.

13. W. Fuchs. Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen. Dusseldorf, 1953.

14. G. Herden. Quantitative Linguistics. London, Butterworths, 1964, p.64 etc,

The Advenced Theory of Language as Choice and Chance. Berlin-Heldelberg-New York, 1966, p.201 etc.

15. Л.Е.Машкина. О статистических методах исследования лексико-грамматической дистрибуции. КД(рукоп.), Минск, 1968.

16. Т.А.Микерина. Некоторые статистические приемы лексико-логического описания функционального стиля (на материале английских текстов по судостроению). АКД, Л., 1967.

17. О.А.Нехай. Статистика и автоматический анализ текста. АКД, Минск, 1968.

18. А.С.Ротарь. Опыт статистического описания современных немецких публицистических текстов (к вопросу о некоторых дивергентных явлениях в газетных текстах ГДР и ФРГ). АКД, Минск, 1971.

19. Л.А.Турыгина. Статистическая интерпретация антиномии «язык и речь». АКД, Ленинград, 1970.

20. Н.В.Смирнов, И.В.Дунин-Барковский. Курс теории вероятностей и математической статистики. И., 1969.

21. П.Ф.Рокицкий. Биологическая статистика. Минск, 1967.

22. В.И.Романовский. Математическая статистика. кн.1.Ташкент, 1961.

23. Л.Хальд. Математическая статистика с техническим приложением. М., 1956.

24. А.Я.Шайкевич. Интервал текста и характер статистических распределений языковых единиц. Сб. «Вопросы лингвостатистики и автоматизации лингвистических работ». Труды ЦНИИПИ, вып. Ш. М., 1969.

25. Г.Крамер. Математические методы статистики. М.,1948.

26. В.И.Романовский. Применение математической статистики в опытном деле. М.-Л., 1947.

27. Л.Н.Большев, Н.В.Смирнов. Таблицы математической статистики. М., 1965.
28. В.Феллер. Введение в теорию вероятностей и ее приложения. т.1, М., 1967.
29. А.Ахабаев. Статистический анализ лексико-морфологической структуры языка казахской публицистики. АКД., Алма-Ата., 1971.

ВЕРОЯТНОСТНЫЕ ЗАКОНЫ РАСПРЕДЕЛЕНИЯ КЛАССОВ СЛОВ КАЗАХСКОГО ТЕКСТА

Труды VII Всесоюзной школы-семинара
«Автоматическое распознавание слуховых образов»
Алма-Ата 18-23 сент. 1972 г. (АРСО – VII).
Изд. «Наука» Каз.ССР Алма-Ата-1973 г.
С. 139-145. (Соавтор – Бектаев К.Б.)

Создание универсальных схем автоматического построения вероятностно-статистических моделей распределения лингвистических единиц в тексте является одной из важных задач прикладной лингвистики. Она необходима, главным образом, для автоматического поиска информации, оптимизации преподавания языков, создания вероятностного базового языка, решения теоретической задачи о вероятностной природе «нормы», в изучении вопросов стилистики, а также при решении многих других задач [1-12 и др.].

В данной работе впервые предпринимается попытка автоматического построения вероятностно-статистической модели распределения классов слов на материале казахских письменных текстов.

В настоящее время лингвостатистика от количественного описания лингвистических явлений, именуемых первым этапом ее развития, переходит на следующий, более сложный этап. Этот этап должен включать построение теории, позволяющей предсказывать количественно те или иные закономерности, получать «уравнение лингвистики» [13, стр. 44].

Таким образом, проблема установления вида распределений единиц текста является одной из наиболее важных задач модельного построения лингвистических явлений. Этому вопросу посвящен ряд исследований, в частности С.Г.Чебанова, Р.М.Фрумкиной, В.К.Калинина, Р.Г.Пиотровского Г.Хердана, Фукса и др. [2, 6, 7, 10, 14-21]. В предлагаемой работе осуществляется проверка гипотез о соответствии статистических распределении предполагаемым теоретическим распределениям (нормальный закон, логарифмически-нормальный, распределение Пуассона и распределение Шарлье в двух типах – А и В).

Статистические распределения существительных, глагола, прилагательных и наречий строились нами на материале выборочно отобранного художественного текста (романа М.Ауэзова «Путь Абая») объёмом 10 тысяч

словоупотреблений и 50 тысяч словоупотреблений публицистического (газетного) текста.

Каждая словоформа выборочно отобранных текстов была отмечена определенной буквой, обозначающей классы слов. Далее, на ЭВМ производилось деление текста на серии (микровыборки) объёмом 25, 50, 100 и т.д. словоупотреблений.

На ЭВМ «Минск-22» по составленной нами программе получен частотный список индексов (индексы определяет классы слов) для каждой серии объёмом 25, 50, 100, 200 и 500 словоупотреблений.

Построенный таким образом дискретный вариационный ряд частот мы используем для проверки соответствия распределений классов слов в текстах теоретическим законам распределения Пуассона и Шарлье типа В. При проверке соответствия распределений классов слов в текстах с другими теоретически законами (нормальным законом, логарифмически-нормальным и законом Шарлье типа А) мы строили непрерывные вариационные ряды путем преобразования дискретных вариационных рядов. Оценки степени близости эмпирического распределения выбранному теоретическому закону распределения производили с помощью неоднократно применяемого в лингвистике критерия согласия χ^2 Пирсона [6,7,10,11,15, 18-22 и др.]. Данная процедура осуществляется следующим образом.

Выдвигается нулевая гипотеза $\mathbf{H_o}$, заключающаяся в том, что эмпирическое распределение выборки соответствует теоретическому распределению генеральной совокупности. Проверку гипотезы $\mathbf{H_0}$ осуществляем по величине вероятности

$$P(\chi^2 \geq \chi_q^2)$$

Если $P(\chi^2 \geq \chi_q^2)$ окажется ниже некоторого заданного наперед уровня значимости, например ниже, чем $q=0,05$, то это будет означать, что вероятность случайных отклонении окажется очень малой величиной. В этом случае мы должны будем признать нашу оценку отклонения эмпирического от теоретического закона неслучайной (ведь случайные явления с очень малой вероятностью следует считать практически невозможными). Это будет указывать на неправдоподобность гипотезы $\mathbf{H_o}$ о несущественности расхождений между эмпирическим и теоретическим распределениями. Если же вероятность достаточно велика (в нашем случае больше 0,05), то нулевую гипотезу о близости обоих видов распределений следует принять.

Значение такой вероятности указывает степень отклонения или приемлимость гипотезы $\mathbf{H_o}$.

Результаты нашего эксперимента и (анализ табличных данных) показывают, что выбор нормировок (объём серии, количество серий объём выборки) существенно влияет на характер распределения изучаемой лингвистической единицы.

Под нормировкой подразумевается следующее: один и тот же текст объёма **N** можно разбить на серии (***k***) объёмом по 100, по 1000, по 2000 и т.д. словоупотреблений, и в соответствии с этим можно выбрать количество

серий ***(n.)***. Причем во всех случаях будет ***N=kn***. В роли серии в лингвостатистике выступает не только определенное количество словоупотреблений, но и «длина» строк, количество страниц и т.д. Следовательно, когда идет речь о распределении той или иной лингвистической единицы, то задача ставится не вообще, а конкретно – с какого объёма серии считать непротиворечивым согласие эмпирического распределения с выбранным теоретическим. Например, мы не можем утверждать, что словоформы распределяются по закону Пуассона, а можем сказать, что при определенных нормировках они описываются законом Пуассона, а при других нормировках, допустим, [10, 29, 23] или, начиная с такого-то объёма серии, описываются определенным вероятностным законом. Далее, всякие сравнения параметров, а также содержательные лингвистические интерпретации можно провести уже на основе этих нормировок [10, 22]. Например при большом параметре закона Пуассона можно пользоваться как нормальным, так и пуассоновым приближением, а при малом – Пуассоновым. Таким образом, нахождение наилучшего способа разбиения текста, безусловно, имеет большое значение в модельном построении лингвостатистики.

Теперь для примера проанализируем распределения существительных.

Существительные, по сравнению с другими частями речи в казахских текстах, являются самыми частыми. Так, например, в тексте романа «Абай жолы» 35,36% слов составляют существительные. В публицистическом тексте существительных 45,19% [24, стр.9]. Поэтому если считать, что выборки удовлетворяет требованиям независимости, то вполне можно предположить, что эмпирические распределения указанных классов слов согласуются с нормальным законом. Проверку данной гипотезы мы осуществляем путем сравнения величин $\mathrm{P}(\chi^2 \geq \chi_q^2)$ для различных комбинаций значений ***N, n, k***.

В таблицах 1 и 2 отражены значения вероятностей $(\chi^2 \geq \chi_q^2)$ соответствия эмпирического распределения существительных с их теоретическими распределениями, т.е. такими, как нормальное, Шарлье типа А и логнормальное, при различных комбинациях значений ***N, n, k.***

Таблица 1. Значеня $\mathrm{P}(\chi^2 \geq \chi_q^2)$ для распределения существительных в художественном тексте

Выборки	**Законы распределения**		
	Нормаль-ный	**Шарлье - А**	**Логнор-мальный**
$\mathrm{B}_{2500}^{100} = 25$	0,660	0,746	0,175
$\mathrm{B}_{5000}^{200} = 25$	0,730	0,548	0,006
$\mathrm{B}_{10000}^{400} = 25$	0,862	0,713	0,0003
$\mathrm{B}_{5000}^{100} = 50$	0,991	0,972	0,368
$\mathrm{B}_{10000}^{200} = 50$	0,625	0,380	0,246
$\mathrm{B}_{10000}^{100} = 100$	0,101	0,193	0,053

Анализ значений вероятностей указанных таблиц показывает, что для всех рассматриваемых объемов серии ***k***, т.е. для ***k***=25, 50, 100, 200, словоупотреблений, объёмов выборов ***N*** (2500, 5000, 10000, 12500, 20000, 25000, 40000, 50000) словоупотреблений и различных количеств серий ***n=N:k***, надежной моделью, описывающей распределение существительных, является нормальной закон распределения. Следующим по мощности, при таком сравнении, оказалось распределение Шарлье типа А. Для существительных газетного текста при объёме серии ***k*≥100** вполне допустимо логнормальное распределение.

Значеня $P(\chi^2 \geq \chi_q^2)$ для распределения существительных в публицистическом тексте

Таблица 2

Выборки	**Законы распределения**		
	Нормальный	**Шарлье - А**	**Логнормальный**
$B_{2500}^{100} = 25$	0,797	0,994	0,999
$B_{5000}^{200} = 25$	0,986	0,936	0,231
$B_{12500}^{500} = 25$	0,740	0,486	0,005
$B_{25000}^{1000} = 25$	0,979	0,880	0,000
$B_{5000}^{100} = 50$	0,687	0,781	0,619
$B_{10000}^{200} = 50$	0,302	0,118	0,064
$B_{25000}^{500} = 50$	0,775	0,398	0,692
$B_{50000}^{1000} = 50$	0,548	0,141	0,0003
$B_{10000}^{100} = 100$	0,626	0,563	0,839
$B_{20000}^{200} = 100$	0,983	0,976	0,750
$B_{50000}^{500} = 100$	0,991	0,990	0,413
$B_{20000}^{100} = 200$	0,310	0,456	0,477
$B_{40000}^{200} = 200$	0,575	0,658	0,677
$B_{50000}^{250} = 200$	0,695	0,699	0,756
$B_{50000}^{1000} = 500$	0,802	0,649	0,809

Литература

1. *Белоногов Г.Г.* О некоторых статистических закономерностях в русской письменной речи. Ж. «Вопросы языкознания», 1962, № 1.

2. *Калинин В.М.* Развитие схемы Пуассона и ее применение для описания статистических свойств речи. АКД., Л., 1964.

3. *Сергал Д.М.* Некотрые уточнения вероятностной модели Ципфа. Сб. «Машинный перевод и прикладная лингвистика». М., 1965, №5.

4. *Зубов А.В., Хотящов Э.Н.* Статистический анализ текста с помощью ЭВМ. Сб. «Энтропия языка и статистика речи». Минск, 1966.

5. *Пиотровский Р.Г.* Моделирование фонологических систем и методы их сравнения. М. –Л., 1966.

6. *Микерина Т.А.* Некоторые статистические приемы лексико-морфологического описания функционального стиля (на материале английских текстов по судостроению). АКД., Л., 1967.

7. *Нехай О.А.* Статистика и автоматический анализ текста. Канд. дисс. Минск, 1968.

8. *Калинина Е.А.* Изучение лексико-статистических закономерностей на основе вероятностной модели. Сб. «Статистика речи». Л., 1968.

9. *Калинин В.М.* Математические аспекты восприятия иноязычного текста . Сб. «Сатистика речи». Л. 1968.

10. *Бектаев К.Б., Лукьяненков К.Ф.* О законах распределения единиц письменной речи. Сб. «Статистика речи и автоматический анализ текста». Л., 1971.

11. *Пиотровский Р.Г. и Турыгина Л.А.* Антиномия «Язык – речь» и статистическая интерпретация нормы языке. Сб. «Статистика речи и автоматический анализ текста». Л., 1971.

12. *Фрумкина Р.М.* Статистические методы изучения лексики. М., 1964.

13. *Добрушин Р.Л.* Математические методы в лингвистике. Сб. «Математическое просвещение». М., 1959, выпуск 6.

14. *Чебанов С.Г.* О подчинении речевых речевых укладов «индоевропейской» группы закону Пуассона. Доклады АНСССР, новая сер., т. 55, 1947. №2.

15. *Фрумкина Р.М.* О законах распределения слов и классов слов. Сб. «Сруктурно – типологические исследования». М., 1962.

16. *Бектаев К.Б., Пиотровский Р.Г.* и др. О законах распределения классов слов, словоформ и словосочетаний. Сб. «Межвузовская конференция по вопросам частотных словарей и автоматизации лингвистических работ». Л., 1966.

17. *Бектаев К.Б., Машкина Л.Е., Микерина Т.А.* Определение констант закона Эсту-Ципфа-Мандельберта в частотных списках словоформ и словосочетаний. Сб. «Энтропия языка и статистика речи». Минск, 1966.

18. *Бектаев К.Б., Зубов А.В., Ковалевич Е.Ф., Машкина Л.Б., Нехай О.А.* К исследованию законов распределения лингвистических единиц. Сб. «Статистика текста». 1, Минск, 1969.

19. *Машина Л.Е.* О статистических методах исследования лексико-грамматической дистрибуции. Канд.дисс., Минск, 1968.

20. *W. Fuchs.* Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen. Dusseldorf, 1953.

21. *G. Herdan.* Quantiative Linguistics. London, Butterworts, 1964, p. 64, etc; The Advanced Theory of Language as Choice and Chance. Berlin – Heidelberg – New-York, 1966, p. 201, etc.

22. *Шайкевич А.Я.* Интервал текста и характер статистических распределений языковых единиц. Сб. «Вопросы лингвостатики и автоматизации лингвистических работ». Труды ЦНИИПИ, вып .Ш. М., 1969.

23. *Феллер В.* Введение в теорию вероятностей и ее прилодения. Том 1. М., 1967.

24. *Ахабаев А.* Статистический анализ лексико-морфологической структуры языка казахской публицистики. АКД, Алма-Ата, 1971.

АКАДЕМИЯ НАУК КАЗАХСКОЙ ССР
ИНСТИТУТ ЯЗЫКОЗНАНИЯ

А.Х.ДЖУБАНОВ

СТАТИСТИЧЕСКОЕ ИССЛЕДОВАНИЕ КАЗАХСКОГО ТЕКСТА С ПРИМЕНЕНИЕМ ЭВМ

(на материале романа М.Ауэзова «Абай жолы»)

10.02.06 – тюркские языки

А В Т О Р Е Ф Е Р А Т

диссертации на соискание ученой
степени кандидата филологических
наук

Диссертация написана на русском языке

Алма-Ата 1973

Работа выполнена в группе «Статистико-лингвистические исследования» Института языкознания автоматизация АН КазССР

Научный руководитель – кандидат физико-математических наук, доцент К.Б.Бектаев.

Научный консультант – академик АН КазССР доктор филологических наук, профессор С.К.Кенесбаев.

Официальные оппоненты:

1. Доктор филологических наук, профессор Р.Г.Пиотровский.
2. Кандидат филологических наук С.С.Татубаев.

Ведущее научное учреждение – Казахский ордена Трудового Красного знамени государственный университет имени С.М.Кирова.

Автореферат разослан «14» мая 1973 г.

Защита состоится «15» июня 1973 г.

на заседании Объединенного Ученого совета АН КазССР г. Алма-Ата, Шевченко, 28.

С диссертацией можно ознакомиться в Центральной научной библиотеке АН КазССР.

Ученый секретарь Совета

Доктор филологических наук М.Т.Дуйсенов

Язык представляет собой сложную многоярусную систему с большим числом параметров изменения. Поэтому вполне естественно применение по отношению к нему методов теории вероятностей, математической статистики, теории информации и др.[1]

При статистическом изучении реального текста необходимо идеализировать его свойства[2] и тем самым получить объект исследования в виде модели, которая явится известным приближением к речевой действительности. Текст, подлежащий моделированию, на первом этапе исследования рассматривается как последовательность независимых друг от друга дискретных единиц, т.е. допускается отсутствие связи между языковыми элементами в потоке речи.

Задача исследователя состоит в том, чтобы обнаружить в данном речевом материале лингвистические реальности, которые не выявляются обычными методами, при этом модель как аналог изучаемого объекта даст возможность распознавать определенные свойства оригинала[3].

Идеи применения вероятностно-статистических, теоретико-информационных методов, а также использование возможностей электронно-вычислительных машин (ЭВМ) для анализа и синтеза языка ныне находят широкую поддержку и со стороны тюркологов[4]. Об этом свидетельствует состоявшаяся в АН КазССР всесоюзная конференция «Статистическое и информационное изучение тюркских языков» (Алма-Ата,1969г.) и функционирование при Институте языкознания АН КазССР группы «Статистико-лингвистические исследования и автоматизация», которая призвана осуществить статолингвистическое описание казахского языка (речи) на всех его уровнях.

Реферируемая работа и являет собой первый опыт применения вероятностно-статистических методов и ЭВМ в исследовании тюркских языков, в частности казахского языка.

Работа состоит из введения, трёх глав, заключения и приложения.

---

[1] Б.Н.Головин. О роли статистики в описании языковых и речевых стилей. Сб. «Частотные словари и автоматическая переработка лингвистических текстов». Минск, 1968.

[2] Е.А.Калинина. Изучение лексико-статистических закономерностей на основе вероятностной модели. Сб. «Статистика речи». Л., 1968, стр.66.

[3] Р.Г.Пиотровский. Моделирование фонологических систем и методы их сравнения. М-Л.,1966,стр.20.

[4] С.К.Кенесбаев, К.Б.Бектаев, Р.Г.Пиотровский. Инженерная лингвистика и тюркология. «Советская тюркология». Баку, 1970, №6; С.К.Кенесбаев, А.Т.Кайдаров, Казахское языкознание за 50 лет. «Вопросы языкознания». 1973, №1; Абай тілі сөздігі. Алматы, 1968; К.Б.Бектаев. Қазақ тілінің кері алфавитті сөздігі (Обратный словарь казахского языка). Алматы 1971; и др. (см. К.Б.Бектаев. «Статистика речи 1957-1972». Библиографический указатель. Алма-Ата, 1972).

Глава I

АВТОМАТИЧЕСКОЕ ПОСТРОЕНИЕ ЧАСТОТНЫХ СПИСКОВ ЛИНГВИСТИЧЕСКИХ ЕДИНИЦ НА МАТЕРИАЛЕ ТЕКСТОВ ТЮРКСКИХ ЯЗЫКОВ

Развитие электронной вычислительной техники показало, что аналогия между мыслительными процессами человека и принципом работы ЭВМ более глубока и многостороння, чем это кажется на первый взгляд. Выявление наиболее типичных общностей человека и машины открывает широкие горизонты моделирования при помощи ЭВМ психо-физиологических функций мозга в соответствии с его разносторонней деятельностью.

Процесс мышления человека протекает естественным путем, что дает основание считать его познаваемым, а деятельность человеческого мозга – алгоритмичной. Поскольку значительная часть человеческого мышления происходит имплицитно, т.е. не всегда осознается человеком, то она ускользает от непосредственного анализа, основанного на самосозерцании. Это значит, что деятельность человеческого мозга на сегодняшнем уровне развития науки можно подразделить на два этапа: алгоритмизируемый (формализуемый этап) и неалгоритмизируемый.

Применение ЭВМ для выполнения формализованных этапов работы мозга человека позволяет резко повысить эффективность процесса научного исследования[6*]. Знание отличительных качеств человеческого мозга и ЭВМ помогает исследователю-лингвисту выявить, какие из этапов работы мозга можно «поручить» ЭВМ, а какие оставить для себя.

Прослеживая процесс переработки информации электронной машиной, нетрудно заметить в ее действиях некоторую аналогию умственной деятельности, а также и ряд отличий. Учитывая их, «целесообразно оставлять машине однотипные, массовые, требующие большого объема работы операции, а сложные эвристические и, самое главное, нестандартные задачи следует оставить человеку».[5] Исходя из этих характеристик своеобразных действий ЭВМ и умственных способностей, мы и строим свою работу.

Наиболее кропотливой работой, выполняемой в процессе автоматического анализа лингвистических единиц, является предмашинное редактирование текста. При этом интересующие нас словесные единицы снабжаются условными пометами-индексами. Такая необходимость возникает из-за ограниченной по сравнению с человеком эвристической способности современной ЭВМ, ибо анализируемые лингвистические единицы текста не всегда могут быть полностью распознаваемыми

---

[5] Э.Я.Якубайтис. Кибернетика и автоматизация труда исследователя. Известия АН Латвийской ССР. 1967.№2.

[6] С.К.Кенесбаев, К.Б.Бектаев, Р.Г.Пиотровский. Указ. Статья (см. сноску 4).

автоматически точно на основе их формальных признаков. Предмашинное редактирование дает возможность осуществить статистическое описание на уровне конфигураций лексических единиц, а также устранить лексико-грамматическую неоднозначность (многозначность и омонимию единиц тексте), являющуюся основной лингвистической трудностью при автоматическом анализе текста. Поскольку значительное количество казахских слов составляет внутри языка омонимические ряды (например, казахские слова «қара», «ақ» могут быть глаголами, прилагательными, существительными), их принадлежность к той или иной категории частей речи выявляется только в контексте.

Первоначально в качестве опыта нами была выработана условная система индексов, состоящая из следующих девяти букв: З-существительное, Г-глагол, С-прилагательное, У-наречие, А-числительное, Е-местоимение, О-междометие, К-служебное слово, Л-подражательное слово, Д-модальное слово.

В дальнейщем, возникла необходимость ввести дополнительные индексы, которые отражали бы и стилистические свойства лексем. С учетом всего этого впоследствии нами была разработана новая, буквенно-цифровая система условных индексов, указывающих на синонимические, омонимические, паронимические и антонимические ряды в составе того или иного текста (преимущественно художественного). С целью получения списков лексических пластов типа архаизмов, неологизмов, модернизмов и т.д. разработаны дополнительные индексы.

Учитывая, что дальнейший этап подготовки текста для ввода в память ЭВМ является довольно трудоемким, желательно редактирование (индексирование) текста производить при решении максимального количества лингвистических задач.

Преобразование текста из его естественной формы в машинную осуществляется при участии человека-оператора. Стало быть, скорость и безошибочность подготовки исходного материала к вводу в ЭВМ зависят от быстродействия и индивидуальных способностей оператора. Кодировка больших массивов проходит, безусловно, не без механических ошибок. Они выявляются только благодаря многократным сверкам контрольной ленты с оригиналом. Весь этот процесс (перфорирование) подобен работе машинистки, что нарушает принцип экономии, связанный с автоматизацией лингвистических исследований.

В своей практике работы на ЭВМ «Минск-22» мы в основном для кодирования текстовой информации пользовались пятипозиционной бумажной перволентой.

Таким образом, предмашинный процесс работы требует от исследователя большого сосредоточения и широкого диапозона знаний, заключающего в себе поготовку как по инженерной, так и по традиционной лингвистике. Тщательное редактирование обрабатываемого текста, а затем его качественное кодирование решают успех статистического анализа языкового материала на ЭВМ.

Аппараты, кодирующие текстовую информацию, вводные и выводные устройства современных ЭВМ основаны на латинской и русской графике. Как известно, подавляющее большинство тюркоязычных народов мира, носителей 19-ти письменных литературных языков[7], проживают в Советском Союзе. Они имеют в своей речи специфические фонемы, которые обозначаются русскими буквами с добавлением диакритических значков. Так как основная цель нашей работы заключается в применении ЭВМ для автоматического анализа текстов тюркских языков вообще, то мы разработали единую транскрипцию специфических букв тюркских языков, которая, главным образом, должна обеспечить ввод текстовой информации в ЭВМ (см.таблицу 1).

Некоторые затруднения в обработке текстов на тюркских языках в системе «человек-машина-человек» не дают повода для скептического отношения к делу и в сравнении с ручной обработкой текста, являются весьма незначительными.

Кроме вышеназванной транскрипции специфических букв, нами разработан также принцип транскрибирования, который допускает ввод в ЭВМ фонетического текста. Исследование фонетической структуры языка способствует определению состава фонем, частотности индивидуальных фонем, а также рядов позиционных вариантов фонем и т.д.

Принципы анализа буквосочетаний, выработанные на графемном уровне, могут быть положены и в основу дешифровки орхоно-енисейских надписей. Исходя из этой мысли, члены группы статолингвистики уже разработали систему индексов, позволяющую уяснить характер дистрибуции графических знаков и установить закономерности их позиционной транскрипции. На сегодняшний день тексты орхонских памятников закодированы (переведены на перфоленты) согласно нашей транскрипции.

Выяснение частот встречаемости и составление списков лингвистических единиц, упорядочение их по интересующему нас признаку (частоте, по алфавиту, по длине и т.д.) играет первостепенную роль в статистическом анализе изучаемого текста. Это дает возможность решения основной задаче по отбору наиболее важных формальных единиц конкретного подъязыка и определению их информационных характеристик, помогает в построении формальной части морфологии, лексики и синтаксиса базового языка (сокращенного варианта естественного языка)[8]. Отсюда возникает необходимость разработки методики составления различных частотных списков. В принципе составление частотных списков ручным способом не является сложным. Но многократное повторение одних и тех же механических (не творческих) операций отнимает много времени и утомляет исследователя, что приводит к всевозможным ошибкам и неточностям.

---

[7] К.М.Мусаев. Алфавиты языков народов СССР.М., 1965, стр.28-34. М.И. Исаев. Сто тридцать равноправных /о языках народов СССР/. Изд-во "Наука", М., 1970, стр. 74-107.

[8] Р.Г. Пиотровский. Базовые языки и отраслевой бинарный перевод. «Язык и общество». Л., 1968.

Таблица 1. **Единая транскрипция специфических букв тюркских языков**

№ / Языки	Азербайджан	Алтайский	Башкирский	Гагаузский	Казахский	Кара-калпакский	Карачево-балкарский	Киргизский	Татарский	Тувинский	Туркменский	Узбекский	Уйгурский	Хакосский	Чувашский	Якутский	Транскрип-ция
1				ä											ă		а’
2	ғ		ғ		ғ	ғ						ғ	ғ	ҕ		ҕ	г’
3															ĕ		е’
4									җ		җ		җ				ж’
5			ҙ														з’
6	к		Ҡ		қ	қ						қ	қ				к’
7		н	ң		ң	ң		ң	ң	ң	ң		ң	ң		н	н’
8	ө	ö	ө	ö	ө	ө		ө	ө	ө	ө		ө	ö		ө	о’
9			с												ç		с’
10		ý		ý	ҰҰ	ý	ý					ý		ÿ	ÿ		у+
11	ҷ													ҷ	ҷ		ч’
12	ә		ә		ә	ә			ә		ә		ә				э’
13	h		ҩ		h	ҳ			h			ҳ	h			h	х’
14	j	j			*i*									*i*		*i*	*i*’
15	ү		ү		ү	ү		ү	ү	ү	ү		ү			ү	у’

На сегодняшний день алгоритмизация задачи построения частотных списков словоформ и др. лингвистических единиц разработана в достаточной мере. Проблемы составления частотных словарей, на ЭВМ явились специальной темой обсуждения ряда всесоюзных конференций[9], освещались на страницах печати и, следовательно, не требуют подробного рассмотрения в данной работе. Однако необходимо отметить, что имеющиеся программы по получению частотных списков словоформ нуждаются в частичном пересмотре в связи с особенностями тюркских языков, которые имеют свои специфические буквы, а также свой порядок расположения букв по алфавиту и введенных нами условных индексов.

В **главе I** нами разработан общий алгоритм составления на ЭВМ частотных списков любой лингвистической единицы. На его основе даны принципиальные алгоритмы построения нижеперечисленных частотных списков:

а) индексов и их сочетаний (индексы могут иметь различный лингвистический смысл, т.е. могут определять лексические, морфологические, или синтаксические и другие признаки слова);

б) букв и буквосочетаний (см.рис.1);

в) интервалов между тождественными буквами (в понятие интервала входит количество букв, пробелов, знаков препинания, заключенных между идентичными буквенными знаками).

Основной отличительной чертой предлагаемых нами алгоритмов программ от имеющихся на сегодняшний день программ по построению на ЭВМ частотных списков букв и буквосочетаний является блок распознавания букв соответствующего тюркского алфавита и блок распознавания слов, не имеющих индекса, от слов с индексом. Алгоритмизация этих задач, конечно, не представляет особой сложности, но необходимость составления таких программ при статистическом исследовании тюркских языков на графемно-фонемном и на морфологическом уровнях бесспорна.

Перечислим возможности программы, составленной для ЭВМ «Минск-22» по алгоритму, приведенному на ***рис.1***.

1. Исходной информацией могут служить тексты, написанные на любом из тюркских языков, тексты, основанные на русской и латинской графике, а также транскрибированный фонетический и др. тексты, имеющие двухсимвольные единицы.

2. Построение следующих алфавитно-частотных списков «графем»:

а) алфавитно-частотный список сплошных сочетаний «графем» с учетом пробела между словами (в каждое последующее сочетание входит предыдущее без его первой «графемы»);

---

[9] Межвузовская конференция по вопросам частотных словарей автоматизации лингвистических работ. Л., 1966; частотные словари и автоматическая переработка лингвистических текстов. Минск, 1968; Всесоюзная конференция по вопросам автоматической переработки текста методами прикладной лингвистики. Кишинев, 1971.

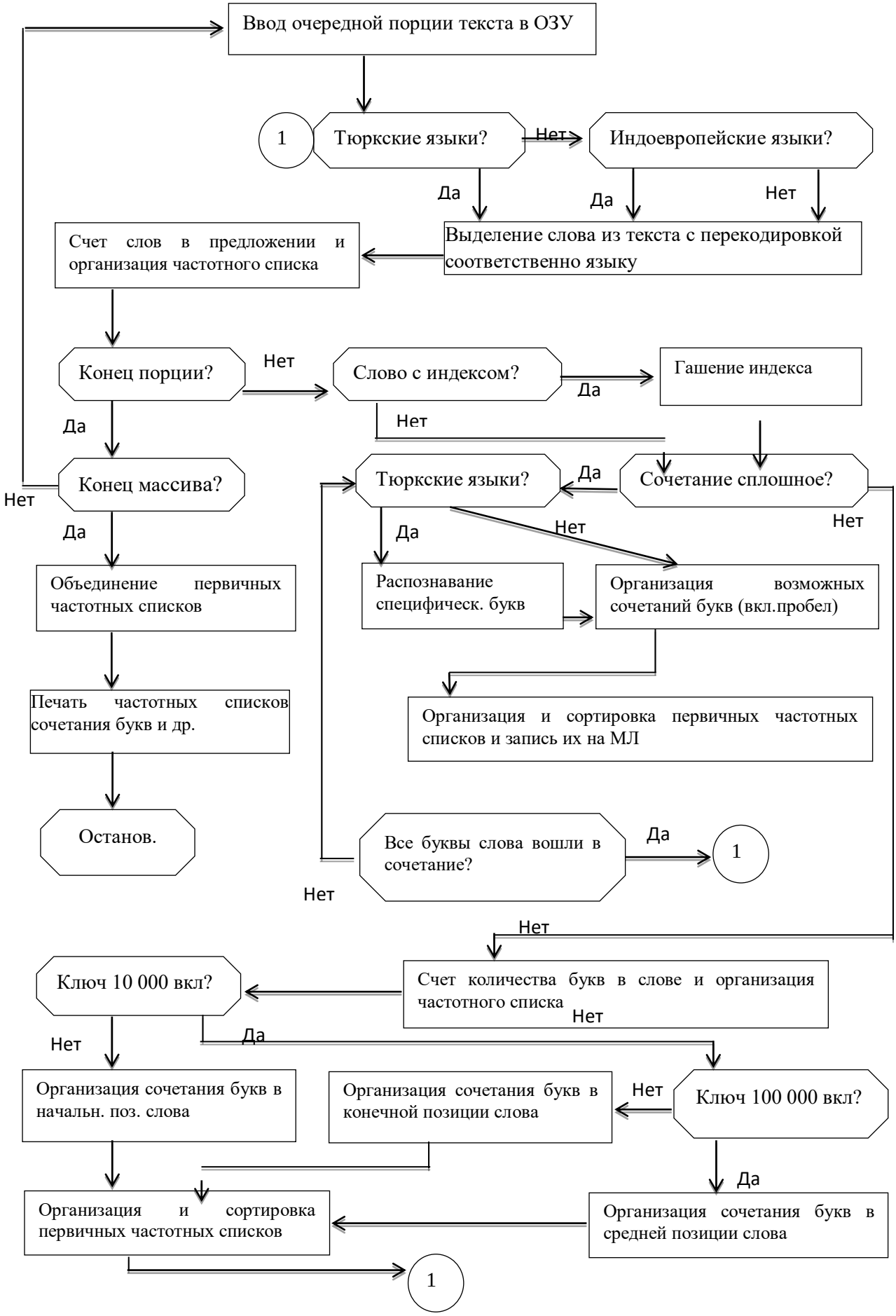


Рис.1. **Алгоритм задачи построения частотного списка букв и буквосочетаний**

б) алфавитно-частотный список сочетаний «графем» в начальной и конечной позиций слова;

в) алфавитно-частотный список сочетании «графем», равноотстоящих от границ (концов) слова, или, как мы условно называем, сочетания букв в средней позиции.

Во всех случаях количество «графем» в сочетании для текстов на тюркских языках равно от 2 до 4-х; для текстов на индоевропейских языках от 2-х до 6-ти.

3. Побочным продуктом программы является частотный список «количество букв в словах» и «количество слов в предложении».

Для того чтобы ЭВМ могла выполнять действия, указанные в алгоритме, необходимо перейти от принципиального алгоритма к алгоритму, записанному языком операции машины, закодировать последний алгоритм – записать его в виде последовательности команд. Именно на этом этапе решения лингвистических задач особенно заметно разделение труда между лингвистами и математиками – программистами. Как показывает опыт работы группы «Статистика речи», наибольший эффект в общении человека и машины может быть достигнут, когда в одном человеке сочеталась бы лингвистическая подготовка с некоторой квалификацией в области программирования[10].

Благодаря реализации составленных нами программ, получены статистические материалы, которые проанализированы в последующих главах.

[10] А.В.Зубов. Переработка текста естественного языка в системе «человек-машина». Сб. «Статистика речи и автоматический анализ текста». Л.,1971,стр.316.

Глава II

ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ТЮРКСКИХ ТЕКСТОВ

В настоящее время лингвостатистика от количественного описания лингвистических явлений, именуемых первым этапом его развития, переходит на следующий, более сложный этап. Этот этап «должен включать построение теории, позволяющей предсказать количественно те или иные закономерности, получать «уравнения лингвистики»[11].

Таким образом, проблема установления вида распределений единиц текста является одной из наиболее важных задач модельного построения лингвистических явлений. Этому вопросу посвящен ряд исследований, в частности С.Г.Чебанова, Р.М.Фрумкиной, В.М.Калинина, Р.Г.Пиотровского, Г.Хердана, В.Фукса, Қ.Б.Бектаева и др.[12]

В данной главе предпринята попытка построения вероятностно-статистической модели распределения классов слов (существительных, глагола, прилагательных и наречий на материале выборочно отобранного художественного текста (романа М.Ауэзова «Абай жолы» объемом 10 тысяч словоупотреблений и публицистического (газетного) текста – 50 тыс. словоупотреблений), приложимости частотной модели Ципфа и функции распределения Вейбулла к текстам на казахском языке (частотные списки романа М.О.Ауэзова «Абай жолы»).

В работе подробно изложена методика построения дискретных и непрерывных вариационных рядов – эмпирических распределений классов слов.

Наряду с такими распределениями, как нормальный, логарифмическо-нормальный законы, закон Пуассона, которые неоднократно применялись в лингвостатистической практике, мы решили для проверки заданных гипотез

---

[11] Р.Л.Добрушин. Математические методы в лингвистике. Сб. «Математическое просвещение», вып.6, М., 1959.

[12] К.Б.Бектаев и К.Ф.Лукяненков. О законах распределения единиц письменной речи. Сб.«Статистика речи и автоматический анализ текста».Л.,1971; К.Б.Бектаев, А.В.Зубов и др. К исследованию законов распределения лингвистических единиц. «Статистика текста». Минск,1969; К.Б.Бектаев, Р.Г.Пиотровский и др. О законах распределения классов слов, словоформ и словосочетаний. «Межвузовская конференция по вопросам частотных словарей и автоматизации лингвистических работ».Л,1966; В.М.Калинин. Развитие схемы Пуассона и ее применение для описания статистических свойств речи. Авторефер.канд.дисс.,Л.,1964; К.Ф. Лукьяненков. Использование схем Пуассона и Гаусса в исследовании распределения лингвистических единиц текста. Сб. «Вопросы лингвостатистики и автоматизации лингвистических работ». Труды ЦНИИПИ, вып.3.М., 1969, стр.5-14; Р.М.Фрумкина. О законах распределения слов и классов слов. Сб. Структурно-типологические исследования». М., 1962, стр. 124-133; С.Г.Чебанов. О подчинении речевых укладов «индоевропейской» группы закону Пуассона. Доклады АН СССР, новая сер.т.55. 1947, №2; W.Fuchs Mathafisch Apalyse yon Sprachelmenten, Sprachen. Dusseldorf, 1953. C.Herdan. Quantinive Linquistics. London, Buttezths. 1964.p. 61 etc. The Advaced Theory of Lanquage as Choice and Chance. Berlin-Heidelberg New York, 1966, p. 201 etc.

использовать в своем эксперименте распределения Шарлье типа А и типа В[13], являющиеся приближенными к вышеназванным распределениям.

Распределение Шарлье типа А является приближенным к нормальному закону распределения и функция его имеет следующий вид:

$$\varphi'(x)=\frac{\Delta x}{\sigma}\cdot\frac{1}{\sqrt{2\pi}}\cdot l^{\frac{(x_i-a)^2}{2\sigma^3}}\cdot\left\{1+\frac{A}{\sigma}\left[\frac{(x_i-a)^3}{\sigma^3}-3\frac{x_i-a}{\sigma}\right]+\frac{E}{24}\cdot\left[\frac{(x_i-a)^4}{\sigma^4}-6\frac{(x_i-a)^2}{\sigma^2}+3\right]\right\}$$

Выражение, стоящее перед фигурной скобкой, не что иное, как нормальный закон распределения. Все параметры распределения Шарлье-А, совпадют с параметрами нормального закона «***a*** – математическое ожидание, ***σ*** – дисперсия, ***A*** и ***E*** – коэффициенты асимметрии и эксцесса).

Распределение Шарлье типа В является приблежённым к распределению Пуассона и функция его имеет следующий вид:

$$\psi'(x)=\frac{a\cdot l^{-a}}{x!}\cdot\left\{1+\frac{\mu_2-a}{a^2}\left[\frac{x(x-1)}{2}-ax+\frac{a^2}{2}\right]\right\}$$

где $\frac{a^x\cdot l^{-a}}{x!}$ – Пуассонова функция распределения, μ_2 – центральный момент второго порядке $(\mu_2=6^2)$.

Близость эмпирических распределений к какому-либо теоретическому распределению должна отражать закономерные соотношения изучаемого признака. Следовательно, возникает необходимость оценки степени близости эмпирического закона теоретическому.

Такую оценку производим с помощью известного в статистике критерия согласия χ^2 Пирсона:

$$\chi_i^2=\sum_{i=1}^{t}-\frac{(n_i^t-N\cdot P_i)^2}{N\cdot P_i}$$

где n_i^t – эмпирическая частота, $N\cdot P_i=n_i^t$ – математическое ожидание (теоретическая частота), P_i – вероятность соответствующая ***i*** – му классу, ***N*** – объем выборки.

Общая схема проведенного эксперимента представляет собой следующую последовательность:

1. Индексирование текста по частям речи.

2. Составление частотного списка индексов (по нашей программе для отдельного случая $B_n^n=K$, где ***K*** – объем серии, ***n*** – количество серии, $N=K\cdot n$ - объем выборки.

3. Построение вариационных рядов и подготовка их к вводу в ЭВМ.

4. Вычисление на ЭВМ значений:

а) среднего арифметического ***a***;

б) среднего квадратичного отклонения ***σ***;

в) коэффициентов асимметрии ***A*** и эксцесса ***E***;

г) теоретических частот $n_i^t=N\cdot P_i$;

д) χ_i^2 для каждой ***i*** – ой группы и их суммы.

[13] В.И.Романовский. Математическая статистика. Ташкент. к.н.I.1961.

5. Объединение (если в этом есть необходимость) частот со значением меньше 5 и введение соответствующей поправки на сумму$\sum_{l=1}^{t} \chi_i^2$.

6. Нахождение числе степеней свободы ***ν=t-l-1***, где ***t*** – число групп эмпирического распределения, ***l*** – число параметров (линейных зависимостей) с учетом того, что для нормального, логнормального распределения и Шарлье–В ***l***=2, для Шарлье типа А ***l***=4 и для распределения Пуассона *l*=1.

7. Нахождение значения интеграла вероятностей $P(\chi^2 \geq \chi_q^2)$ по вычисленному χ^2 и числа степеней свободы ***ν*** (из статистических таблиц) и сведение их в таблицу.

Анализ полученных табличных данных показывает, что выбор нормировок (объем серии, количество серий и обьем выборки) существенно влияет на характер распределения изучаемой лингвистической единицы. Следовательно, когда идет речь о распределении той или иной лингвистической единицы, задачу надо ставить конкретно – с какого объема серии считать непротиворечивым согласие эмпирического распределения о выбранным теоретическим. Всякие сравнения параметров, а также содержательные лингвистических интерпретаций можно провести уже на основе этих нормировок[14].

Существительные и глоголы в казахском языке (так же, как и в других языках) являются наиболее часто встречающимися классами слов. Так, например, в тексте романе 35% слов составляют существительные и 31,44% – глаголы. В публицистическом же тексте[15] существительные – 45,19%а глаголы – 21,86%. Поэтому если считать, что результативность выборки удовлетворяет требованиям независимости, то вполне можно предположить, что эмпирические распределения существительных и глаголов казахского языка согласуются с нормальным законом распределения.

Путем сравнения величин $P(\chi^2 \geq \chi_q^2)$ для различных комбинаций значений ***N, n, k*** мы провели проверку нулевой гипотезы (***H₀***) о нормальности распределения указанных классов слов. Анализ значений вероятностей показывает, что для всех рассматриваемых нами объемов серии К (К=25, 50, 100, 200 словоупотреблений), объемов выборок (2500, 5000, 10000, 12500, 20000, 25000, 40000, 50000 словоупотреблений) и различных количеств серий ***n=N:K***, надежной моделью, описывающей распределение существительных, является нормальный закон. Следующим по надежности оказалось распределение Шарлье типа А. Для существительных газетного текста при объеме серии ***K≥100*** вполне допустимо логнормальное

---

[14] К.Ф.Лукьяненков. Использование схем Пуассона и Гаусса в исследовании распределения лингвистических единиц текста. Сб. «Вопросы лингвостатистики и автоматизации лингвистических работ». Труды ЦНИИПИ, вып. 3, М.,1969, стр.5-14; А.Я.Шайкевич. Интервал текста и характер статистических распределений языковых единиц. Там же, стр.15-22, В.Фаллер. Введение в теорию вероятностей и ее приложения. Т.I, М., 1967:

[15] А.А.Ахабаев. Статистический анализ лексико-морфологической структуры языка кзахской публицистики. Автораф.канд.дисс. Алма-Ата, 1971, стр. 9.

распределение. При других значениях «***K***» гипотеза H_0 не во всех случаях подтверждается и для романа и для публицистического текста.

Анализ степени согласия эмпирического распределения казахских глаголов в художественном тексте с теоретическим распределением показывает, что нормальный закон может служить математической моделью глагольных классов слов. Правда, для тех случаев, когда $K \leq 100$, наилучшее соответствие имеет распределение Шарлье типа А. В публицистических же текстах, начиная с $K \geq 100$ для различных $n \geq 100$ и $N \geq 10.000$, наилучшее согласие остается за нормальным распределением. Это может дать основание предположить, что распределение глаголов казахского языка, так же как и распределение существительных, подчиняется нормальному закону (по крайней мере для указанных нормировок). Оценка существенности различия между эмпирическими распределениями существительного, глагола, и теоретическими – Пуассона и Шарлье типа В показала явное их несоответствие.

Исследование на согласие эмпирических распределений прилагательных художественного и публицистического текстов с рассматриваемыми теоретическими распределениями дает следующие результаты. Для значений объемов внутри серийной выборки $K \geq 25$, для различных $n \geq 100$ и $N \geq 2500$ художественного текста и для $K \geq 50$, $n \geq 100$ и $N \geq 5000$ публицистического текста распределение прилагательных в подавляющем большинстве случаев согласуется с законами Шарлье типа А, В и Пуассона. При такой ситуации предпочтение можно отдать распределению Пуассона в силу его удобства и изученности. Однако, начиная с $K \geq 200$, $n \geq 100$ и $N \geq 20000$, нормальное распределение также хорошо согласуется с эмпирическим распределением прилагательных.

Наречия распределены в тексте в большинстве случаев по законам Пуассона и Шарлье типа В. Удобным для применения, конечно, является распределение Пуассона, но при $K > 200$ предпочтительным может быть нормальный закон, т.к. в наших экспериментах хорошее согласие эмпирического распределение наречий с упомянутым теоретическим распределением замечается при нормировке: $K \geq 100$, $n \geq 100$ *и* $N \geq 10000$.

Таким образом, здесь предпринята попытка найти наилучший способ разбивки текста, при котором эмпирические распределения казахских классов слов могут достаточно хорошо соглосоваться с определенным теоретическим законом распределения. Результаты подобных исследований, на наш взгляд, имеют большое значение при изучении вопросов моделирования лингвистических фактов и явлений.

Как известно, структура частотных словарей любого языка обладает в целом общую закономерность. Однако некоторые параметры этой закономерности в зависимости от строя того или иного языка или же от стилистического своеобразия и объема текста колеблются в определенной степени. Изменения подобного рода служат для исследователя характерными признаками, присущими описываемым текстам. Исходя из такой

предпосылки, для описания распределения вероятностей появления словоформ в частотном словаре, мы, согласно методике Г.Г.Белоногова

использовали формулу Вайбулла[16], которая имеет следующий вид:

$$f_i^{*\grave{o}\mathring{a}\grave{i}\eth} = 1 - \mathring{a}^{ci^k},$$

где $f_i^{теор}$ – теоретическая накопленная относительная частота, i – порядковый номер словоформ в частотном списке, ***с*** и ***к*** – параметры формулы Вейбулла.

Указанная формула уже применялась для вычисления теоретической относительной накопленной частоты словоформ и трехсловных сочетаний в текстах на русском, румынском, немецком, англииском и др. языках[17].

Определив значение параметров ***С*** и ***К*** для каждого номера слов методом наименьших квадратов, можно вычислить соответствующее им значение. Все вычисления производились на ЭВМ[18].

Математическую оценку близости вычисленных теоретических частот с эмпирическими частотами словоформ мы попытались дать с использованием «λ – критерия» А.Н.Колмогорова, который отличается своей простотой[19].

Проведенные нами подсчеты, касающиеся частотных словарей для каждой из пяти книг и романа в целом (М.Ауэзова «Абай жолы»), приводятся в следующей ***таблице 2***:

Таблица 2

Параметры	Словари				
	I	II	III	IV	V
Д	0,0107	0,0103	0,0106	0,0106	0,020
λ	0,612	0,641	0,634	0,673	1,962
Р (λ)	0,864	0,864	0,864	0,711	0,002

Статистические данные, указанные в таблице, вычислены для средний части частотных списков (с 55-го порядкового номера и до номера, охватывающего абсолютную частоту, равную пяти). Значения вероятностей Р (λ), вычисленные без исключения высокочастотной и низкочастотной зон словаря, получились очень малыми, примерно 0,001 и 0,002, что, конечно, опровергает принятую гипотезу о том, что случайная величина, являющаяся эмпирической относительной накопленной частотой, имеет закон распределения, заданный функцией распределения Вейбулла. Для романа в целом значение Р(λ) в указанной таблица также имеет малую величину. Это

[16] Г.Г.Белоногов. О некоторых статистических закономерностях в русской письменной речи. «Вопросы языкознания», 1962, №1.

[17] К.Ф.Лукьяненков. Лексикостатистическое описание английского научно-технического текста с помощью ЭВМ.Автореф.канд. дисс..Минск,1969; Н.Г.Маткаш. Лексика и морфология молдавской публицистики в сравнении о лексикой и морфологией других дако-романских функциональных стилей.Автореф.канд.дисс.,Л., –1967.

[18] А.В.Зубов., З.Н.Хотяшов. Статистический анализ текста с помощью ЭВМ. Сб. «Энтропия языка и статистика речи». Минск,1968.

[19] Е.С._Вентцель. «Теория вероятностей» М.,1964,стр.149-158.

свидетельствует о том,что в зависимости от изменения объема текста меняется средняя зона словаря.

Таким образом, формула Вейбулла может служить моделью, описывающей вероятность появления словоформ в частотном словаре только для средней ее части (т.е. когда $i > 50$).

Статолингвистам широко известен так называемый закон Ципфа, который устанавливает зависимость между частотой слова и его номером:

$$F=N\cdot K(i+\rho)^{-\gamma}$$

где, k, γ, ρ – параметры закона (в таком виде и литературе формула носит название «Формула Эсту-Ципфа-Мандельброта»).

Как известно, подробное исследование распределении закона ципфа на материалах индекса Хэнли, словаря языка Пушкина и др. было проведено Р.М.Фрумкиной[20]. Ее исследования выяснили характер изменения параметра γ в разных зонах частотного списка, для разных текстов и различных их объемов.

Придерживаясь ее матодики, мы провели исследование закона ципфа на материале пяти частотных словарей романа «Абай жолы». Полученные нами результаты подтверждают выводы, сделанные Р.М.Фрумкиной относительно параметров γ, k, ρ. Внутри одного произведения (романа) автора замечаются различные значения , γ, k, ρ. Причем близкие по объему тексты имеют и близкие значения γ (см ***таблица 3***) .

Таблица 3

частотные словари	Объемы текстов	конеч № при $F>4$	ρ	γ	k
1-кн.	105788	3270	4,770	0,942	0,092
2-кн.	124398	3872	5,490	0,937	0,088
3-кн.	112727	3570	4,760	0,924	0,079
4-кн.	123053	4027	6,380	0,930	0,086
Роман в целом	465966	9614	7,835	1,017	0,146

В ряде работ была показана целесообразности вычисления γ, k, ρ относительно участков частотного списка[21]. С этой целью мы произвели условное деление частотных списков романа на 19 участков (интервалов) . Для всех участков на ЭВМ были зачислены параметры γ и $lg(NK)$ через коэффициенты прямолинейной регрессии. Составленная нами программа учитывает любые деления частотного списка на зоны. Для этого достаточно задать необходимое число пар начальных и конечных порядковых номеров зон. Анализируя полученные данные, можно сделать заключение, что параметр γ в пределах одного текста (жанра) в разных зонах частотного списка принимает различные значения, близкие к 1. Некоторое отклонения происходят за счет начальной и конечной зон. Среднее арифметическое

[20] *Р.М.Фрумкина.* Статические методы научения лексики. М. «Наука» 1964. стр.36; так же: К вопросу о том называемом «Законе Ципфа». «Вопросы языкознания»,1961, №2, стр.117-122.

[21] *К.Б.Бектаев, Л.Е.Машкина, Т.А.Империна.* Определение констант закона Эсту-Ципфа-Мандельброта в частотных списках словоформ и словосочетаний. Сб. «Энтропия языка и статистика речи», Минск,1966; Д.М.Сегал. Некоторые уточнения вероятностной модели Ципфа.Сб. «Машинный перевод и прикладная лингвистика». 1961.№5 и др.

значение γ (по 19 зонам), вычисленное для частотного списка романа в целом, получилось равным 0,913. Значения параметра «***К***» подверждают ранее высказанные утверждения Кондона к Ципфа о том, что «***К***» приблизительно равно 0,1. Колебания значений «***К***» по зонам схожи с колебаниями параметра γ.

Таким образом, все «утверждения и выводы, касающиеся закона Ципфа, сделанные по частотным спискам текстов на индоевропейских языках, в основном сохраняются и для частотных списков тюркских текстов, в частности казахского художественного текста.

С целью выявить существование связи между рангом и частотой словоформ в пяти частотных списках романа мы построили опытные линии регрессии и, предполагая на их основе корреляционную связь прямолинейной, нашли уравнение прямой регрессии методом наименьших квадратов. Для того чтобы установить, одинакова ли сила связи между исследуемыми величинами по всей длине частотного списка или она меняется на его различных участках, были вычислены по нашей программе необходимые параметры для расчета уравнения регрессии.

Судя по значениям коэффициентов корреляции, можно утверждать о наличии прямолинейной корреляционнной связи, обратной зависимости между рангом и частотой во всех частотных списках романа. Сравнение силы связи и отклонений теоретических прямых по отдельным зонам свидетельствует о довольно сильной зависимости в средней части списка по отнашению к начальному и конечному участкам списка. Сильную зависимость между рангом и частотой (в пяти частотных списках) можно видеть по среднему арифметическому значению коэффицента и по оценке надежности его вычисления,приведенных в ***таблице 4***.

Таблица 4

частотные словари	I ч.с.	II ч.с.	III ч.с.	IV ч.с.	V ч.с. (роман)
r σ_r	-0,966 0,000326	-0,969 0,000276	-0,967 0,000325	-0,971 0,000284	-0,943 0,000144

Глава III

СТАТИСТИЧЕСКАЯ СТРУКТУРА ТЕКСТА РОМАНА М.АУЭЗОВА «АБАЙ ЖОЛЫ»

В филологической науке немало монографических работ, посвященных изучению стилистики, языка видных деятелей казахской литературы. Однако имеющиеся в них элементы статистического анализа носят сугубо иллюстративный характер, в основным приемом исследования остается метод традиционной лингвистики.

Выбрав объектом исследования художественный текст произведения отдельного писателя и основываясь на частотных характеристиках словоформ, классов слов и букв, мы попытались описать статистическую структуру языка романа М.Ауэзова «Абай жолы». Полученные частотные списки дали возможность раскрыть ряд лингвистических закономерностей, действующих в художественной речи индивидуального автора.

Остановимся на некоторых соотношениях между текстом романа и его словником и на отдельных характеристиках частых и редких словоформ. В результате анализа частотных словарей четырех книг романа и объединенного словаря (романа в целом) мы выяснили следующее: текст романа состоит из 465966 словоупотреблений, 1-ая книга содержит 105788, 2-ая книга – 124498, 3-ья книга – 112727 и 4-ая книга – 123053, словоупотребления. Число различных словоформ в каждом из четырех частотных словарей соответственно равно: 22642, 26418, 26530, 27447, а в объединенном списке – 61824 словоформы.

Рассматривая отношение «длины» словника (***l***) к «длине» текста (***N***), (т.е. *l*:*N*), можно в какой-то степени судить о богатстве лексикона автора. Например, чем ближе значение указанного отношения к единице, тем больше в тексте различных словоформ. Отношения, вычисленные на материале пяти словарей романа, имеют соответственно такие значения: 0,214, 0,212, 0,235, 0,223 и по роману в целом – 0,133. Отсюда следует существенный вывод: если предстоит сравнить богатства лексикона двух или более авторов, то необходимо исходить из одинакового объема текстов ($N_1 = N_2 = N$).

Изучение статистических структур текстов различных языков с целью установления общих черт было впервые предложено известным английским статистиком Дж.Юлом в связи с его попытками установления авторства анонимных текстов[22]. В своих исследованиях он пришел к выводу, что сравнение статистических структур различных языков возможно только при условии равенства длины текстов, установил чрезвычайную схожесть их структур. В результате анализа количественных соотношений текста романа и его словников мы пришли к следующим выводам:

[22] G.V. Yule. The statistical study of literary, vocabulary. Cambridge,1944

1. Незначительный процент словоформ (по отношению к длине словника), находящихся в высокочастотной зоне словаря, покрывает довольно большой процент текста. Например, для 4-х книг романа в среднем 755 наиболее частых словоформ, т.е. около 3% всех словоформ, а также 842 (1,4%) частых словоформ романа в целом, соответственно покрывают 50% текста.

2. Чем меньше частота словоформ, тем больше количество словоформ с данной частотой. Например, количество словоформ с частотой 1, по отношению к длине словника, составляет 60%, с частотой 2 – 13%, в частотном списке словоформ, имеющих частоту 30 и более насчитывают всего 3% и т.д. Эти данные являются прямым доказательством того, что богатство словарного запаса автора действительно заключается в редких словах[23]. В этом можно убедиться, анализируя и проценты покрываемости текста группами словоформ с одинаковой частотой.

Поскольку с увеличением частоты количество словоформ в группах уменьшается, то соответственно уменьшается процент покрываемости текста ими.

3. Установлено, что среднее количество информации, которое приходится на одну единицу частотного списка, т.е. на одну словоформу казахского художественного текста (романа), равно 12 дв.ед. это значительно больше в сравнении с количеством информации, вычисленной для индоевропейских языков. Данный факт говорит о том, что казахская словоформа располагает сравнительно большими возможностями для связи с другими словоформами.

В данной главе комментируется сто (100) высокочастотных словоформ с точки зрения лингвистической закономерности. Эту часть словаря условно называем зоной высокой частоты. Так, соединительные союзы: ***да//де***, ***мен//пен***, ***және***; вопросительная частица ***ма***; связки типа: ***соң***, ***сияқты, қарай, үшін***, ***ал***, ***ғана***, ***бірақ***, ***тағы***, ***ғой*** и др. служебные слова в собственно ауэзовской стилистике обнаруживают способность к организации наиболее излюбленных оборотов писателя. Глаголы с основами: ***еді***, ***екен***, ***етіп***, а также ***деп***, ***деді***, ***деген***, образованные от ***«е»*** и «***де»***, вернее, все деепричастные формы авктивно используются в синтаксических конструкциях, типа **глагол+глагол** или **имя+глагол**.

Известная четверка казахских вспомогательных глаголов: ***отыр, жатыр, тұр, жүр***, в рассматриваемой зоне фигурирует как ***отыр***, ***отырып***, ***отырған***, ***жатқан***, ***жатыр***; ***тұр***, ***тұрған***, ***тұрып***, ***жүрген***. Как отметил проф. Х.К.Жубанов: «... Эти четыре глагола ... могут непосредственно к своей основе принимать личные окончания I и II лица так же, как и имена, в своем корневом виде имеют форму чистого презенса; в то же время, оставаясь основой, т.е. без всяких наращений, они могут иметь форму двух лиц: II лица императива и III лица презенса и т.д.»[24], что и послужило

[23] *Р.М. Фрумкина.* Статистические методы изучения лексики. Изд-во «Наука», М., 1964, стр.51.

[24] *Х.К.Жубанов.* Исследования по казахскому языку. Алма-ата, «Наука», 1966., стр.76.

причиной высокочастотности в языке писателя, кроме них в первой сотне высокочастотных словоформ оказались: причастия на –***тын***, ***болатын***; на -***ған***// -***ген***, ***келген***, ***болған***; деепричастия на -***ып***//-***іп***//-***п***: ***алып***, ***айтып***, ***келіп***, ***беріп***, ***боп***, ***көп***; на -***е***: ***келе***; глаголы: условного наклонения на –***са***: ***болса***; изъявительного наклонения на -***ды***//-***ді***, -***ты***//-***ті***: ***болды***, ***берді***, ***келді***, ***айтты***, ***кетті***.

Высокочастотными явились также слова модального значения, как-то: ***жоқ*** (нет), ***бар*** (есть), ***аз*** (мало), ***көп*** (много). Не выражая конкретно ни качества, ни количества предмета, они выполняют функции как именных частей речи, так и глагола, в силу чего реализуется в речи намного чаще, чем другие категории лексико-семантических образований.

Широко представлены в перечне высокочастотных словоформ разряды местоимений, прилагательные: а) качественные – ***үлкен*** (большой), ***қатты*** (твердый), ***қалың*** (толстый), ***жақсы*** (хороший), ***қара*** (черный), ***жас*** (молодой); б) относительные – ***жаңағы*** (недавний), ***соңғы*** (последний), образованные от основ: ***жаңа*** (недавно), ***соң*** (после), которым присуще свойство отражать отличительные признаки предметов или явлений.

В исследуемом романе имена прилагательные использованы писателем 38531 раз, что составляет 8,3% от всего словоупотребления – 465966. Без учета повторяемости различные словоформы прилагательных встречаются 4773 раза. Это 7,7% всех словоформ в романе, насчитывающем 61824 этих единиц. Их различные структурные формы представлены в ***таблице 5***. Примечательно, что процентное содержание корневых и производных форм прилагательных в тексте приблизительно одинаковое.

Таблица 5

формы прилагательных	**процент от длины словника прилагательных**	**процент от всего объема прилагательных**
Корневые	14,9	47,7
Производные	66,4	45,4
Из парных слов	5,50	1,60
Из слитных слов	5,40	1,80
Степени сравнения	5,10	2,70
Интенсивные	1,20	0,60
С нетюркскими	0,70	0,40
аффиксами	0,80	0,20
Заимствованные		
Всего прилагательных	4773 (100%)	38531 (100%)

Подобное равновесие в употреблении прилагательных исходит не из тождества их словника (использованных словоформ). Напротив, в авторском употреблении корневые словоформы составляют 14,9% словника,тогда как производным формам отводится 66,4%. Это означает,что писатель в целом по роману больше использовал качественные прилагательные, выражающие признаки предметов и явлений в своей морфемной структуре, нежели относительнее, указывающие на то или иное свойство описываемого предмета только исходя из характеристик аналогичных явлений. Причиной

высокочастотности в языке писателя, кроме них в первой сотне высокочастотных словоформ оказались: причастия на –***тын***, ***болатын***; на -***ған***// -***ген***, ***келген***, ***болған***; деепричастия на -***ып***//-***іп***//-***п***: ***алып***, ***айтып***, ***келіп***, ***беріп***, ***боп***, ***көп***; на -***е***: ***келе***; глаголы: условного наклонения на –***са***: ***болса***; изъявительного наклонения на -***ды***//-***ді***, -***ты***//-***ті***: ***болды***, ***берді***, ***келді***, ***айтты***, ***кетті***.

Среди качественных прилагательных, выражающих различные цвета и масти животных, с наибольшей частотой использовано прилагательное <u>қара</u> (черный) – 644 раза. Как известно, цвета черный и белый в казахском языке обозначают не только окраску предмета, но и такие абстрактные понятия, как виновный и невиновный, злой, добрый и т.д. Из остальных цветов часто употреблены: ***көк*** (синий) – 226, ***қызыл*** (красный) – 200, ***қоңыр*** (коричневый) – 171, ***сары*** (желтый) – 145 и т.д.

Не менее интересен подбор прилагательных для характеристики психо-физиологических свойств. В этом разряде с наибольшей частотой употребляется прилагательное ***қатал*** (суровый) – 102. Почти антонимичное ему слово ***момын*** (тихий, безвредный) встречается 96 раз, ***сыпайы*** (вежливый) – 92, ***адал*** (честный) –78, ***нәзік*** (нежный) – 69 раз. С одинаковой частотой употреблены прилагательные с положительной окраской; ***әділ*** (справедливый) – 45, ***сергек*** (чуткий) – 45, ***батыл*** (смелый) – 44. Равномерно распределяется также целый ряд прилагательных с отрицательной окраской.

Качественные прилагательные семантически переплетаются двумя способами: как антонимы и синонимы. Причем одни синонимические ряды вступают с другими в антонимические отношения. Разумеется, по частоте употребляемости подобные противопоставления не могут соблюдать какого-либо равновесия. В диссертации дается весьма условно сгруппированный перечень антонимии синонимических рядов.

В романе использованы и прилагательные, образованные от казахских основ, с прибавлением к ним нетюркских (иранских) аффиксов типа: ***алда-м-<u>паз</u>, боқта-м-<u>паз</u>, өнер-<u>паз</u>, ал-ым-<u>паз</u>, шал-ым-<u>паз</u>, жаз-ым-<u>паз</u>, ресм-<u>и</u>, әскер-<u>и</u>, дүние-<u>үй</u>, шарх-<u>и</u>, <u>бей</u>-қам, <u>бей</u>-маза, <u>би</u>-шара, <u>би</u>-тарап, <u>бәд</u>-бахыт, <u>на</u>-разы, <u>на</u>-риза*** и т.д. Они составляют незначительный процент как в словнике, так и в самом тексте. Нередко встречаются и слова иноструктурного типа, заимствованные из арабо-иранских языков, они служат для окраски речи отдельных персонажей, как, например, ***аһли***, ***ғибади***, ***тахлиди***, ***шағиғазиз***, ***фак,*** а также их фонетические варианты, ***ғазіз***, ***әзіз***, ***пак***, ***пәк*** и т.д. Обращает внимание также малый процент употребляемости (см. ***табл. 5***) интенсивной формы прилагательных, т.е. с усилительным слогом в препозиции.

Известный лингвистический интерес представляют принципы подхода писателя к формообразующим аффиксам степеней сравнения прилагательных. Так, аффикс – ***қай*** употребляется автором всего 45 раз. Из них в трех случаях он прибавляется к трем прилагательным: ***шым+<u>қай</u>, сұр-<u>қай</u>, тырбы(қ)+<u>қай</u>***, а в остальных 42 случаях – только к прилагательному ***қоңыр*** (темный) в виде ***қоңыр+<u>қай</u>.*** Высокая частота данного слова

объясняется широкими семантическими возможностями обозначать все темные цвета, а также служить для обозначения различных эмоций.

Привлекают внимание такие синонимические понятия, как ***сұргылт***, ***сұрқыл***, ***сұрқай***, образованные от прилагательного ***сұр*** (серый). Малую частоту прилагательного ***сұрқыл*** и ***сұрқай*** надо рассматривать, по-видимому, как результат стилистической особенности писателя, прибегшего к словотворчеству путем необычной грамматикализации.

Для более ясного представления о статистике аффиксов прилагательных в стилистике писателя мы сопоставили данные II книги романа «Абай жолы» (124398 словоупотреблений) с публицистическим текстом современного казахского языка (150000).

По нашим подсчетам в указанной книге романа встречается 11467 (9,2%) прилагательных с учетом повторяемости. Аналогичный объем публицистического текста содержит 18534 (12,4%) прилагательных. Из всего числа прилагательных, использованных М.Ауэзовым, корневые слова составляют 51,3%, а аффиксальные образования – 48,7%. Судя по этим отношениям, можно утверждать, что в романе соблюдено приблизительное равенство между этими двумя видами имен прилагательных. В публицистике же аффиксальная форма прилагательных заметно превалирует над корневыми (см.табл.6). Такое различие между двумя жанрами, по-видимому, объясняется тем что художественному тексту слова, обозначающие яркие краски естественных тонов, тогда как газетные и журнальные тексты характеризуются более сдержанными выражениями в смысле красочности.

Подсчет аффиксов осуществлялся нами по трем разновидностям: употребление аффиксов отыменных прилагательных, от глагольных прилагательных и малопродуктивных древних аффиксов. Процентные соотношения вышеуказанных классов аффиксальных прилагательных отражены в ***таблице 6***.

Таблица *6*

классы аффиксальных прилагательных	роман				публицистика			
	колич. аффик.	%	частота аффик.	%	колич аффик.	%	частота аффик.	%
имя+аффикс	38	46,9	4987	89,8	39	42,9	11832	91,2
глагол+аффикс	21	25,9	365	6,6	27	29,7	768	5,9
основа+древний аффикс	22	27,2	235	3,6	25	27,4	375	2,9
Итого:	81	100	5587	100	91	100	12975	100

В процентном отношении нет больших расхождений в употреблении аффиксов рассматриваемых трех классов. Однако не все аффиксы одинако употреблены в романе и публицистике. Равномерность распределения аффиксов отсутствует и внутри каждого текста. Так, 12 из 38 аффиксов отыменных прилагательных в романе составляют 88% употребляемости. Только один формант ***-ғы*** использован писателем 1006 раз, что составляет 20,2% всей частоты. К тому же мягкие варианты указанных аффиксов составляют 34,8%, а твердые варианты – 65,2%. Причем, 11% из 34,8% принадлежит только форманту ***-ғі***.

Аналогично в тексте публицистики 82,2% употребляемости падает только на 12 аффиксов из 39. Такие аффиксы,как ***-ғы/-гі, -сыз/-сіз, -ды//-ді, -ты//-ті, -лы//-лі, -дай//-дей***, как в романе, так и в публицистике, находятся в зоне высокой частоты. В отличие от них аффиксы ***-дық//-дік, -тық//-тік, -лық//-лік***, имеющие высокую активность в публицистике, не являются таковыми в романе. В тексте публицистики твердые и мягкие варианты аффиксов отыменных прилагательных встречаются примерно в одинаковой степени (51,6% и 48,4% соответственно).

В количественном отношении аффиксы отглагольных прилагательных встречаются в равной мере как в романе, так и в публицистике. Но в отношении частоты аффиксов существует некоторый разнобой. Так, в романе отсутствуют форманты ***-ңқы//-ыңқы, -ба//-бе//-пе, -палы//-інді, -шек, -ақ//-кіш***. В свою очередь в текстах публицистики отсутствуют аффиксы ***-мды//-ымды, -нды***, которые в романе встречаются относительно часто.

Формант ***-ыс*** использован во II кн. романа с частотой более чем 50 раз, тогда как в публицистическом тексте зарегистрирован всего один раз. В процентном отношении только два аффикса ***-с, -ыс*** покрывают 30,4% аффиксального образования. В научной грамматике казахского языка форманты ***-с, -ыс//-іс*** отнесены к разряду непродуктивных аффиксов[25]. Данные романа отражают иную картину. Писатель образует при помощи их целый ряд прилагательных типа: ***тозбас, аралас, қимас, қажымас, айнымас, жасымас, аямас*** и т.д. То есть ***-с, -ыс//-іс***, хотя с малой частотой, но встречаются довольно активно, снабжая множество глагольных основ. Это означает, что писатель, обладающий богатым словарным фондом, не прибегает к частому употреблению одних и тех же словесных величии, а многократно использует известные форманты с различными корневыми морфемами.

Примечательно, что в романе из 30-ти разновидных формантов, образующих отглагольное прилагательное, только шесть ***(-ы//-і, -імді, -с//-ыс, -лы***) составляют 70,7% всего употребления. Остальные 24 форманта покрывают лишь 29,3%.

Наиболее продуктивные 8 аффиксов в языке публицистики покрывают 90,1% всей накопленной частоты отглагольных прилагательных. Наибольший процент употребляемости принадлежит форманту ***-ма***, который встречается 287 раз, что составляет 37,2% всех остальных форм. При том формант ***-ма*** в тексте публицистики оформляет лишь два слова – ***құрама*** (273 раза) и ***тоқыма*** (14). Твердые варианты аффиксов отглагольных прилагательных составляют в тексте романа – 57,5%, в публицистике – 68%. Однако здесь следует учитывать высокую частотность форманта – ***-ма*** в последнем тексте, ибо без него процент твердых аффиксов равнялся бы 30,8%. Мягкие варианты аффиксов соответственно составляют 42,5% в романе и 32% в публицистике.

[25] Қазақ тілінің грамматикасы. I, морфология, Алматы,1967, 80-бет.

На примере этих данных можно проследить наибольшую последовательность сингармонизации слов у писателя, нежели в языке публицистики. С другой стороны, подобная черта романа подчеркивает обязательность внутренней гармонии для эстетического целого, каковым является произведение литературы, чего, однако, нельзя требовать от разрозненных публицистических текстов.

В научной грамматике казахского языка учтено более 30 так называемых древних (или репродуктивных) аффиксов, которые в современном языке являются как бы закрепленными за отдельными корневыми морфемами. В нашем исследовании установлено, что частота подобных аффиксов мала. Однако реализация древних формантов имеет свои особенности в языках художественно-индивидуализированного произведения и публицистики.

Нашему анализу подвергнуты все графемы (ГФ) казахской письменной речи на основе трех стилистически разных источников, а именно: текста романа «Абай жолы» (художественный текст – ХТ), данных двухтомного толкового словаря (ТС) казахского языка и текста научно-технической литературы (НТЛ). Выявление частот ГФ производилось как без учета их позиций в словах, так и с их учетом, а именно: встречаемость ГФ в начальной и конечной позициях. Гласные ГФ подразделены на широкие и узкие, а согласные ГФ – на сонорные, глухие и звонкие. Процентные соотношения ГФ как по разделам, так и по стилям отражены в ***таблицах 7-8***.

Таблица 7

Позиции	**Тексты**	**Согласные (%)**				
		сонорные	**глухие**	**звонкие**	**итого**	**Объём текста (в графемах)**
Началь-ная	ХТ	4,70	36,02	30,15	70,87	51290
	ТС	7,21	44,59	23,80	75,60	17585
	НТЛ	10,80	37,10	29,80	77,70	19467
Конечная	ХТ	38,42	19,61	2,05	60,08	51173
	ТС	18,90	21,36	3,20	43,46	17542
	НТЛ	37,00	13,20	2,79	53,09	19444
Без учета позиции	ХТ	22,60	19,95	14,31	56,86	280813
	ТС	21,20	23,20	11,00	55,40	116317
	НТЛ	24,60	17,71	14,50	56,81	130388

Таблица 8

Позиции	**Тексты**	**Гласные (%)**			
		широкие	**узкие**	**итого**	**Объём текста (в графемах)**
Началь-ная	ХТ	23,73	5,40	29,13	51290
	ТС	17,30	7,10	24,40	17585
	НТЛ	19,40	2,90	22,30	19467
Конечная	ХТ	17,98	21,94	39,92	51173
	ТС	12,06	44,48	56,54	17542
	НТЛ	20,41	26,50	46,91	19444
Без учета	ХТ	25,93	17,21	43,14	280813

позиции	ТС	24,40	20,20	46,60	116317
	НТЛ	25,27	17,92	43,19	130388

Анализируя ГФ трех источников без учета позиций как гласных, так и согласных в слове, можно сделать некоторые предварительные выводы о том, что по своей статистической характеристике тексты романа и НТЛ более близки между собой по сравнению с текстом ТС.

При рассмотрении ГФ с точки зрения занимаемых ими позиций в словах жанровые различия трех источников обнаруживаются более явственно. Так, если сопоставить разделы гласных ГФ, то видим, что наибольший процент встречаемости во всех трех источниках имеют широкие гласные. Соотношения процентных содержаний их с узкими гласными в среднем составляют 4:1. Рассмотренные по жанрам, эти соотношения выглядят следующим образом: по роману – 4:1, по словарю – 2:1, по НТЛ – 6:1.

При рассмотрении согласных ГФ в начальной позиции можно ужверждать, что для казахских слов, начинающихся с согласных, характерна большая активность шумных согласных. Глухие же согласные в начале слов встречаются чаще, чем звонкие.

Во всех лексикографических трудах исходной формой казахского глагола принято считать инфинитив который оформляется аффиксом «у» в конце слова. Данное положение наиболее явственно подтверждается и статистической характеристикой конечных гласных. Так, в словаре конечные гласные активизировались за счет узких гласных (44,5%) точнее, за счет гласного «у», на который падает 35% встречаемости всех конечных гласных.

При рассмотрении конечной позиции выявляется, что сонорные согласные, составляющие весьма незначительный процент в начальной позиции, в конце слов составляют более половины частот от общего числа, т.е. по роману – 38,4%, по НТЛ – 37%, звонкие же согласные в конце слов казахской речи оказались не характерными, встречаемость их очень мала – 2,05%.

Более независимую роль от изменений позиций играют глухие согласные. Анализ частот отдельных графем по позициям показывает, что в исследуемых источниках встречаемость их не одинакова. Причем во всех трех источниках некоторые ГФ (в любой позиции слова) встречаются чаще остальных. Например, среди гласных ***а, е-ё, ы, і*** (называемые универсальними фонемами,) являются самыми частыми как без учёта занимаемых позиций в слове, так и с учетом. Притом высокочастотность данных ГФ характерна для всех трех источников. Суммарный процент встречаемости этих ГФ выглядит таким образом: без учета позиций в романе – 34,7% в словаре – 28,7%, в НТЛ – 32% тогда как встречаемость всех гласных в трех источниках составляет 43,1%, 44,6%, 43,2% соответственно. В позиционном отношении также подтверждается продуктивность универсальных фонем.

Обобщая данные относительно частотности отдельных графем, можно сказать, что для казахской письменной речи, равно как и для звуковой речи, характерны сонорные и смычно-взрывные ГФ.

В приложении к диссертации даны тексты программ с инструкцией по их использованию. Фрагменты частотных списков, полученных на ЭВМ «Минск–22» и другие статистические материалы.

Автор настоящей работы выражает большую благодарность А.В.Зубову за его полезные советы в деле разработки алгоритмов программ.

Основные положения диссертации отражены в следующих работах:

1. Автоматизация некоторых лингвистических процессов. «Вестник–АН КазССР», 1968, №9; 6 стр. (совместно с А.В. Зубовым).

2. Автоматическое построение частотных словарей (прямого и обратного) «Вестник АН КазССР», 1970, №3; 6 стр. (совместно с А.В.Зубовым).

3. Информационное и статистическое изучение тюркских языков. «Известия АН КазССР», 1971, №3; 3стр. (совместно с А.В.Зубовым и С.Мирзабековым).

4. Исследование статистической структуры тюркского текста// «Автоматическая переработка текста методами прикладной лингвистики». Материалы всесоюзной конференции 6-8 октября 1971г. Кишинев, 1971; 2 стр. (совместно с К.Б.Бектаевым и др.).

5. К вопросу автоматизации лингвистических исследовании «Материалы первой научной конференции молодых ученых АН КазССР». Издательство «Наука», Алма-Ата, 1968; 2 стр.

6. О некоторых путях вскрытия вероятностно-статистических закономерностей в современном казахском языке. «Статистическое и информационное изучение тюркских языков» (СИИТЯ). Тезисы докладов и сообщений. Алма-Ата, –1969; 2 стр.

7. О частотности формообразовательных аффиксов существительных казахского языка. СИИТЯ. Тезисы докладов и сообщений. Алма-Ата, 1969; 2 стр. (совместно с А.Ахабаевым и др.).

8. Статистика аффиксов имен прилагательных. СИИТЯ. Тезисы докладов и сообщений. Алма-Ата,1969; 4 стр.(совместно с У. Шойбековым).

9. Структурно-статистическая модель описания грамматических категорий тюркских текстов. «Структурно-математические методы моделирования языка». Тезисы докладов всесоюзной конференции. Киев, 1970; 2 стр. (совместно с К.Б.Бектаевым и др.).

10. Сөздік қорымыз қанша? (Каков запас слов?). «Білім және еңбек», №3, Алматы, 1971; 4 стр. (совместно с С.Мирзабековым).

11. Универсальная программа получения на ЭВМ «Минск-22» алфавитно-частотного списка сочетаемости графем. «Материалы второй научной конференции молодых ученых АН КазССР» . Алма-Ата, 1970; 2 стр.

12. Частотные словари и информационные характеристики языка. «Тезисы докладов научно-теоретической конференции профессорско-

преподавательского состава института, посвященной 50-летию Великой Октябрьской социалистической революции». Чимкент, 1967; 2 стр. (совместно с К.Б.Бектаевым и др.).

КИБЕРНЕТИКА И ЛИНГВИСТИКА

Материалы третьей научно-теоретической
конференции молодых ученых АН Каз. ССР
(исследования на стыке наук). Часть II.
Алма-Ата 1974, С. 245-247. (Соавтор - Б.Хасанов)

При всей специфике и сложности объектов общественных наук применимость к ним идей и средств кибернетики, а также математических методов вполне обоснована, ибо в гуманитарных исследованиях не должно быть субъективизма, умозрительности, расплывчатости в описании фактов и явлений.

В своих трудах, посвященных вопросам применимости кибернетики в различных гуманитарных науках, Б.В.Бирюков, Е.С.Геллер и др. довольно убедительно раскрывают сущность этого применения, основываясь на диалектико-материалистических принципах единства количества и качества, формального и содержательного подходов.

Среди гуманитарных наук на первое место в смысле применения кибернетики следует поставить науку экономическую, но кибернетика показала себя и в других областях гуманитарных знаний: в психологии, праве, теории и практике обучения человека, лингвистике, истории и др.

Как известно, математическая лингвистика развивалась вместе с работами по автоматизации перевода с одного языка на другой, которые развернулись с конца 40-х годов во многих странах мира, и именно идеи кибернетики во многом питают новые области науки о языке – структурную и математическую лингвистику.

Работы в области машинного перевода и опыты по автоматизации операции над языком показали неудовлетворительность – с точки зрения формализаций – существующих грамматик естественных языков, содержащих неясные и даже противоречащие друг другу формулировки и пр.

Таким образом, задачи машинного перевода требуют нового подхода к языку – аналогичного подходу к искусственным знаковым системам, что не только необходимо для решения задач автоматической обработки текстов на естественных языках, но в более широкой, теоретической перспективе ведет к познанию глубинных закономерностей человеческого мышления и общения между людьми.

Применение математических методов в исследовании фактов языка наряду с прикладным значением имеет и огромное значение для собственно лингвистических изысканий. Стремление к точности, к объективности

выводов – характерная черта любой науки, в том числе и одной из наиболее точных гуманитарных наук – науки о языке.

В настоящее время в индоевропейском языкознании довольно широко стали применяться методы теории вероятностей и математической статистики с использованием электронно-вычислительных машин (ЭВМ).

В области тюркского языкознания они не были столь распространены, если не считать отдельных исследований, основанных на количественном подходе к языковым явлениям, суть которых сводилась в основном к установлению процентного содержания лингвистических данных на том или ином уровне языка. Необходимо отметить, что реальные основания для количественного изучения тюркских языков и их свойств заложены в самой их языковой структуре.

Среди тюркоязычных республик только в Казахстане в АН КазССР должное внимание было уделено изучению казахского языка математическими методами с применением ЭВМ. Как доказательство сказанному можно отметить функционирующую при Институте языкознания АН КазССР группу «Статолингвистические исследования и автоматизация», которой руководит кандидат физико-математических наук, доцент К.Б.Бектаев.

Мы в данном докладе хотим остановиться на одном из статистических исследований проведенных нами на материале текста романа – эпопеи М.Ауэзова «Абай жолы» и публицистического текста, а именно на вопросе установления вида распределений единиц текста, который является одной из наиболее важных задач модельного построения лингвистических явлений. В гуманитарных науках моделирование – задача сложная и касается она в большинстве случаев больших систем (диффузных, плохо организованных систем). В таком случае модель описывает лишь отдельные, наиболее интересные явления, протекающие в сложной диффузной системе.

В своем эксперименте мы вначале построили статистическое (эмпирическое) распределение существительных, глагола, прилагательных и наречий. Как вероятностные модели приняли нормальный закон, логарифмически-нормальный, распределение Пуассона и распределение Шарлье в двух типах – *A* и *B*. Так как близость эмпирических распределений к какому-либо теоретическому распределению отражает закономерные соотношения изучаемого признака, то возникает необходимость оценки степени близости эмпирического распределения выбранному теоретическому закону распределения. Такую оценку производили с помощью критерия согласия $\boldsymbol{X^2}$ Пирсона.

Эксперимент проводился в такой последовательности:

1. Индексирование текста по частям речи.
2. Составление на ЭВМ частотного словаря частей речи.
3. Построение вариационных рядов и подготовка их к вводу в ЭВМ.
4. Вычисление на ЭВМ значений:
   а) среднего арифметического – а ;
   б) среднего квадратического отклонения – б *;*

в) коэффициентов асимметрии А и эксцесса Е ;

г) соответствующих законам теоретических частот $n_i^T = NPi$, по соответствующим формулам; д) X^2 для каждой *i*-ой группы и их суммы.

5) Объединение частот со значением меньше пяти (если в этом есть необходимость) и введение соответствующей поправки на сумму $\sum_t x_i^2$.

6. Нахождение числа степеней свободы *У=t-l-1*, где *t* число групп эмпирического распределения, число параметров (линейных зависимостей).

7. Нахождение значения интеграла вероятностей $P(x^2 \geq x^2 q)$ вычисленному $\boldsymbol{X^2}$ и числа степеней свобода *Y* (из статистических таблиц) и сведение их в таблицу.

Анализ полученных табличных данных показывает, что выбор нормировок (объем серии, количество серий и обьем выборки) существенно влияет на характер распределения изучаемой лингвистической единицы.

Следовательно, когда идет речь о распредедении той или иной ливгвистической едининицы задачу надо ставить конкретно с какого объема серии считать непротиворечивым согласие эмпирического распределения с выбранным теоретическим. Всякие сравнения параметров, а также содержательные лингвистические интерпретации можно провести уже на основе этих нормировок. Результаты подобных исследований, на наш взгляд, имеют большое значение при изучении вопросов моделирования лингвистических фактов и явлений.

В заключении следует отметить, что наше общество имеет большие возможности использования теоретических и технических средств кибернетики в науках, изучающих общественные явления и человека, но при этом, конечно, и не может быть речи о какой-либо подмене кибернетикой общественных наук и сужении области их исследований.

ИНФОРМАЦИОННО-СТАТИСТИЧЕСКАЯ ТИПОЛОГИЯ ТЮРКСКОГО ТЕКСТА

Отделение литературы и языка АН СССР,
Отделение истории АН СССР, Советский комитет
тюркологов при ОЛЯ АН СССР, АН Каз. ССР.
Всесоюзная тюркологическая конференция
27-29 сентября 1976 г. «Советская тюркология
и развитие тюркских языков в СССР»
Тезисы докладов и сообщений. С. 305-307
(Соавторы: К.Б.Бектаев, Р.Г.Ппиотровский)

1. Для типологических исследований чрезвычайно актуальна выработка унифицированных методов выявления универсалий и типологчческой специфики. В докладе представлена первая попытка создания единообразной информационно-статистической технологии,

которую можно было бы использовать для типологических исследований текста.

Поскольку в прямом наблюдении нам дан только текст, представляется целесообразным вести типологические исследования с помощью предлагаемых единообразных приемов лингвистики текста, ориентируясь на письменный текст. Это позволяет, с одной стороны, применять более или менее единообразную процедуру выделения таких элементов, как словоформа, силлабограф, графема, а с другой – использовать в качестве метаязыка язык теории информации и статистики.

2. Исходя из этих соображений было осуществлено на ЭВМ информационно-статистическое исследование большой совокупности тюркских и ивдоевропейских текстов, насчитывающих несколько миллионов словоупотреблений (всего более +10 тыс. страниц).

Использование в ходе этого исследования ряда лингво-математических критериев позволило выявить два типа информационно-статистических характеристик:

а) речевые универсалии,

б) типологические текстовые индексы.

В качестве речевых универсалий выступают:

1) величина вероятностной и информационной покрываемости наугад взятого текста фиксированными массивами наиболее частых слов (для тюркских языков вероятностная покрываемость тысячью наиболее частых слов лежит в интервале 0,63÷0,86, для индоевропейских – в промежутке 0,64÷0,89, соответственно интервалы информационной покрываемости составляют для тюркских языков 0,59÷0,78 для индоевропейских – 0,53÷0,77);

2) избыточность текста (в усредненном «общеязыковом» тексте она колеблется в пределах 70÷80 %*);*

3) нормальность распределения средних длин словоформ в языках мира;

4) неподчиняемость доминантных (в том числе терминологических слов нормальному, логнормальному, пуассоновскому законам распределения.

Таблица 1

Языки	Среднее количество информации на словоформу $I_{сф}$	Доли информации, приходящейся на лексико-грамматические связи в тексте $h_{лг}$ %	Доля информации, передаваемой словоформой $h_{вс}$ %	Коэффициент синтетичности
Казахский	12,47	18,0	82,0	2,80 : 7,90
Русский	11,52	22,5	77,5	1,28
Немецкий	10,42	29,2	70,8	0,05 : 1,70
Румынский	10,32	27,1	72,9	1,02 : 1,35
Французский	9,37	23,5	76,5	0 : 0,82
Английский	9,34	35,0	65,0	0 : 0,5

3. Статистико-информационное исследование тюркских языков в сопоставлении с индоевропейским материалом обнаружило следующие типологические особенности (индексы) тюркского текста:

1) высокая информационная насыщенность словоформы ($I_{сф}$ $Ḥ_{вс}$) и сравнительно слабая её зависимость от контекста ($Ḥ_{лг}$) – см. табл. 1;

2) низкая средняя повторяемость словоформы ($Ꝼ_{сф}$); в агглютинативном казахском языке $Ꝼ_{сф}$ принимает значения в интервале 3,94÷5,49; а во флективно-аналитических, французском и английских языках она попадает в интервал 9,80÷18,04;

3) более высокая по сравнению с индоевропейскими языками частотность употребления знаменательных слов (словоформ);

4) высокое значение параметра ρ в формуле закона Ципфа-Мандельброта (табл. 1);

5) особый характер распределения силлабографов в тюркском слове: в казахском, татарском и турецком языках это распределение характеризуется пологим графиком с правосторонней скошенностью, для флективно-аналитического английского языка, обладающего элементами аморфности, характерен более симметричный островершинный график.

К ВОПРОСУ О ГРАФЕМНОЙ СТАТИСТИКЕ КАЗАХСКОГО ТЕКСТА

АН КазССР Институт языкознания
«Вопросы казахской фонетики и фонологии».
Алма-Ата-1979. С. 79-86.

Письменная речь, как и звуковая, – важнейший языковой материал, требующий практического и теоретического изучения. Характерные особенности письменной речи обнаруживаются в различных литературных жанрах в качественных и количественных соотношениях письменного текста.

В данной статье сделана попытка статистического изучения казахских письменных текстов на графемном уровне, ибо графема − простейшая наглядная единица, изображающая речь средствами письма. Материал исследований может служить основой для фонологических и фонетических работ, анализа морфемной структуры языка и синтагматического рассмотрения языковых единиц. Для облегчения статистических подсчетов мы разработали алгоритм задач получения на электронно-вычислительной машине частотного списка букв и буквосочетаний и составили частотные списки букв с учетом их положения в слове [1]. Материалом для частотных списков послужили графемы (ГФ) казахской письменной речи на основе текстов трех разных по жанру источников: романа М.О.Ауэзова «Абай жолы»; двухтомного толкового словаря казахского языка [2]; математической литературы [3].

Из романа взято 51290 словоупотреблений, что составило 280812 ГФ, из словаря − 17585 словоупотреблений (116317 ГФ), из математической литературы − 19467 словоупотреблений (130388 ГФ) (Объемы выборок мы брали на основании имеющихся частотных списков букв, полученных при испытании машинной программы). Из-за различных объемов выборок числовые характеристики определяли на основании их относительной частотности.

Графемы делили на группы букв, обозначающих гласные (***а, ә, е-ё, и, о, ө, у, ұ, ү, ы, і, э, ю, я)*** и согласные звуки (***б, в, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, ф, х, h , ц, ч, ш, щ***). ГФ гласных в свою очередь разграничивали на широких (***а, ә, е, о, ө, э, я***) и узких (***и, у, ұ, ы, і, ю***); ГФ согласных − на подгруппы сонорных (***л, м, н, ң, р, й***), глухих (***к, қ, п, с, т, ф, х, h, ч, ш, щ***) и звонких (***б, в, г, ғ, д, ж, з, ц***). Далее подсчитывали относительную частотность отдельных ГФ и их сумму в подгруппах текстов отдельных жанров (табл. 1- 3).

Относительную частотность ГФ подсчитывали в трех случаях: в начале слова, в конце слова, во всех положениях.

Рассмотрим некоторые характерные в количественном отношении особенности распределения графем казахского письменного текста.

ГФ согласных звуков составляют 57% текста, ГФ гласных – 43%.

В исследуемых текстах во всех позициях ГФ широких гласных преобладает над ГФ узких гласных на 7-9%; ГФ сонорных согласных употребляются чаще, чем глухих, а последние чаще, чем ГФ звонких (см. ***табл. 1***).

Анализ частотности ГФ в начале слова показал, что более 70% казахских слов в рассматриваемых текстах начинаются с ГФ согласных. Среди них более употребительны ГФ глухих согласных (в среднем 40%), а также звонких (около 29%). ГФ сонорных в начале слова оказалось лишь от 5 до 11 %.

Таблица 1. Распределение ГФ согласных в исследуемых текстах, %

Положение ГФ	**Текст**	**ГФ согласных**				
		сонорные	**глухие**	**звонкие**	**всего**	**%**
Начальное	Роман	4,70	36,02	30,15	70,87	71%
	Словарь	7,21	44,59	23,80	75,60	76%
	Математика	10,80	37,10	29,80	77,70	78%
Конечное	Роман	38,42	19,61	2,05	60,08	60%
	Словарь	18,90	21,36	3,20	43,46	43%
	Математика	37,0	13,20	2,79	53,09	53%
Во всех положениях	Роман	22,60	19,95	14,31	56,86	57%
	Словарь	21,20	23,20	11,00	55,40	55%
	Математика	24,60	17,71	14,50	56,81	57%

Таблица 2. Распределение ГФ согласных в исследуемых текстах, %

Положение	**Текст**	**ГФ согласных**			
		широкие	**узкие**	**всего**	**%**
Начальное	Роман	23,73	5,40	29,13	29%

	Словарь	17,03	7,10	24,40	24%
	Математика	19,40	2,90	22,30	22%
Конечное	Роман	17,98	21,94	39,92	40%
	Словарь	12,06	44,48	56,54	57%
	Математика	20,41	26,50	46,91	47%
Во всех положениях	Роман	25,93	17,21	43,14	43%
	Словарь	24,40	20,20	44,60	45%
	Математика	25,27	17,92	43,19	43%

Подробное статистическое исследование конечных, букв и буквосочетаний казахских слов убеждает в том, что конец слова является ареной действия не только фонетических, синтагматических, но и морфологических факторов.

В данном случае мы затрагиваем только статистику последней буквы слова. Как видно из ***таблиц 1*** и ***2***, более половины всех слов (53-60%) в художественной и математической литературе оканчивается на ГФ согласных звуков. Причем чаще других встречаются ГФ сонорных (37-38%), менее частотны ГФ глухих (13-20%). В редких случаях слова оканчиваются на ГФ звонких согласных (не более 3%).

ГФ согласных в конце слов в художественных и математических текстах встречаются в меньшей половине (40-47%) всех слов. ГФ широких и узких согласных в конечной позиции встречаются приблизительно в одинаковом количестве (см. ***табл. 2***).

При сопоставлении относительной частотности расположения графем в начале и конце слов по тексту толкового словаря с данными текстов из художественной и математической литературы можно отметить некоторые различия (см. «Вопросы казахской фонетики и фонологии». Алма-Ата-1979. С. 82-83, ***табл. 3***). В словаре все производные и непроизводные основы имеют только изначальную форму, в других двух источниках зафиксированы словоформы с учетом их повторения в тексте. Подобное состояние слов (в одном случае статическое, в другом − динамическое) в известной мере влияет и на процентное соотношение графем в тексте. Напрашивается вывод о том, что нельзя распространять статистические данные полученные по текстам словарей, на другие виды текстов.

Статистика группы ГФ гласных и согласных без учета их положения в слове подтверждает выводы А.А.Исенгельдиной о том, что жанровые различия не оказывают существенного влияния на относительную частотность больших групп фонем, какими являются группы гласных и согласных, взятые в целом [4]. Данный факт свидетелстьвует о существовании прямых соответствий между графемами и фонемами казахского языка, что доказывает тезис о фонематическом характере письменной речи.

Рассматривая графемы в позиционном отношении в слове, можно высказать несколько иное предположение: группы ГФ гласных и согласных и их подгруппы в каждом из рассматриваемых текстов несколько отличаются друг от друга в процентной содержании.

Анализ частот отдельных графем показывает (см. «Вопросы казахской фонетики и фонологии». Алма-Ата-1979. С. 82-83, ***табл. 3***), что встречаемость ГФ гласных и ГФ согласных в исследуемых источниках неодинакова, некоторые графемы встречаются чаще остальных, например, ***а, е-ё, ы, і*** − так называемые универсальные гласные. Их высокочастотность характерна для всех трех источников. Суммарный процент встречаемости этих четырех графем без учета их положения в слове в романе 34,7, в словаре − 27,7, в математическом тексте − 32 . Эти данные можно считать довольно высокими, так как частотность всех ГФ гласных, вместе взятых, составляет по трем источникам соответственно 43,14, 44,6 и 43,2%.

В начале слова указанные универсальные графемы также обладают большей частотностью. Так, если встречаемость всех ГФ гласных по трем источникам в начале слова 26%, то половина этого количества относится к ГФ гласных ***а, е- ё, ы, і***.

Вывод о том, что в статистическом плане не всегда возможно сопоставить данные словаря и текстовых источников подтверждается и на примере универсальных графем. В сплошных текстах эти ГФ гласных более активны именно в конечном положении слова. Например, по нашим подсчетам в романе 40% слов оканчиваются на ГФ гласных, из них 38,5% слов имеют в конце только графемы ***а, е, ы, і***. В математическом тексте конечные ГФ гласных составляют 47%, из них универсальным графемам отводится 44%.

После универсальных ГФ гласных по частотности особенно в сплошном подсчете графем, следует ГФ гласных ***о, у, и***. Замечается большая частотность в начале слове ГФ гласнчх ***ә, о, ө***.

Подобный анализ ГФ согласных трех источников с учетом их положения в слове и без него дает иные результаты. Универсальность высокочастотных ГФ согласных здесь не замечается. При сопоставлении ГФ согласных по положению в слове обнаружено, что некоторые сравнительно частые в начальной позиции ГФ одного источника являются таковыми и в двух других, К ним относятся: ***б, к, қ, ж, т, с, д, м***.

В конечном положении в слове высокочастотную зону образуют всего три ГФ согласных: ***н, р, к***. В сплошном же подсчете графем в словах одинаково частотны в трех источниках ГФ согласных ***н, т, р, д, л, с***.

Обобщая данные, можно сказать, что для казахского письменного текста, как и для казахской речи, характерны по относительной частотности ГФ сонорных и ГФ смычно-взрывных звуков.

Ряд ГФ согласных и гласных звуков типа ***в, х, ч, h ,щ, ц, э, ю, я*** , а также ***ь-ъ*** проникли в казахский язык через заимствованные (русские, арабо-иранские) слова. Эти графемы во всех рассмотренных случаях имеют наименьшую частоту, и в процентном отношении каждая из них не превышает 1%.

Литература

1. *Джубанов А.Х.* Задача получения на ЭВМ частотных списков лингвистических единиц. -В кн.: Статистика казахского текста. Алма-Ата, 1973.

2. Қазақ тілінің түсіндірме сөздігі. 1, 2 т. Алма-Ата, 1959, 1961 ж.

3. *Кочетков Е.С., Кочеткова Е.С.* Алгебра және элементарлық функциялар. Алма-Ата. 1967, 1-2 бөлім.

4. *Исенгельдина А.А.* Факторы определяющие относительную частотность фонем. -В кн.: Статистика казахского текста, Алма-ата, 1973, с.659.

О ХОДЕ ПОДГОТОВКИ К СОСТАВЛЕНИЮ СЛОВАРЯ ЯЗЫКА М.О.АУЭЗОВА (Функции ЭВМ)

Мин.Прос.Каз.ССР Чим-кий пединститут.
Материалы семинара «Статистическая оптимизация преподавания языков и инженерная лингвистика».
Чимкент. 1980. – С. 193-195.

Прошло более 10-ти лет с тех пор, как языковедами АН КазССР был издан «Словарь языка Абая», который отражает словарное богатство и специфику употребления слов великим казахским поэтом Абаем Кунанбаевым.

Сейчас, когда подходит к завершению выпуск «Толкового словаря казахского языка» в 10-ти томах, в Институте языкознания АН КазССР начались подготовительные работы по составлению словаря языка произведений М.О.Ауэзова – выдающегося писателя советского Казахстана, лауреата Ленинской и Государственной премий.

Словарь будет составляться по вновь издаваемому собранию сочинений М.О.Ауэзова, По предварительным подсчетам в 20-ти томах произведений М.О.Ауэзова более 2500000 словоупотреблений. Такой большой объем требует иного подхода составителей словаря к методике накопления информации.

В традиционном методе такие операции, как выписывание карточек-цитат, упорядочение их по алфавиту, установление частоты употребления слов и т.д. выполнялись вручную. Эти трудоемкие процессы можно облегчить, используя возможности ЭВМ.

При этом предварительно должен быть решен ряд вопросов, от которых зависит успех применения ЭВМ..

Прежде всего необходимо уяснить технологию подготовки и контроля вводимой в ЭВМ информации, а также выбор комплекса технических средств.

На наш взгляд, наиболее надежным и высокопроизводительным средством для первоначального накопления информации является

комплексное устройство подготовки данных на МЛ ЕС–9002 и алфавитно-цифровых дисплеях ЕС–7906.

Подготовка информации на МЛ (ЕС-9002) сочетает в себе преимущества перфокарт (дискретность и надежность) и перфолент (плавающий формат, отсутствие возможности нарушения последовательности перфокарт). Кроме того, при использовании МЛ достигается большая экономия носителя информации (МЛ), а также машинного времени, т.к. скорость ввода с МЛ в 25–30 раз выше, чем с других носителей информации.

Очень ответственным звеном является контроль подготовленной текстовой информации. Для этих целей мы пользуемся алфавитно-цифровыми дисплеями ЕС-7906, которые позволяют одновременно с контролем на экране вести корректировку текста и записывать исправленную информацию на магнитный диск ЭВМ.

Какую же информацию ожидают от ЭВМ составители словаря языка произведений М.О.Ауэзова?

Во-первых, по старой традиции, они хотят иметь картотечную базу по произведениям автора, т.е. 2,5 млн. карточек-цитат, во-вторых, получить алфавитно-частотный словарь на основе анализа текстов объемом 2,5 млн. словоупотреблений.

Это невозможно осуществить за короткий срок. Достаточно указать на то, что решение этой задачи равносильно перепечатыванию на ЭВМ указанных 20 томов книг в 2,5 млн. экземплярах.

Таким образом, на ЭВМ будут получены следующие материалы:

1) распределительный (по томам) алфавитно-частотный словарь словоупотреблений, с указанием соответствующего тома и страницы, а также частоты употребления слов как по отдельным томам, так и в целом;

2) все цитаты, соответствующие словоупотреблениям в словаре, будут храниться на магнитном диске ЭВМ.

Составители словаря, имея распределительный словарь словоупотреблений, без особого труда могут отыскивать в книге соответствующие цитаты.

Если же составитель сочтет нужным обязательно иметь перед собой карточки-цитаты, где встречаются данные словоупотребления, то по своевременному заказу могут пользоваться напечатенными на АЦПУ карточками-цитатами. При наличии связи с ЭВМ, составители словаря могут прочесть интересующие их цитаты также через экран дисплеев.

На сегодняшний день группа сотрудников Института языкознания АН КазССР проводит необходимые эксперименты по подготовке текстовой информации на устройстве ЕС-9002 и алгоритмизации всей указанной задачи.

ПРЕОБРАЗОВАНИЕ С ПОМОЩЬЮ ЭВМ КАЗАХСКИХ ТЕКСТОВ В ФОНЕТИЧЕСКУЮ ЗАПИСЬ

Мин.Прос.Каз.ССР Чим-кий пединститут.
Материалы семинара «Статистическая оптимизация преподавания языков и инженерная лингвистика».
Чимкент. 1980. – С. 317-318. (Соав.: А.Джунисбеков).

Фонетическая запись (транскрипция) казахского текста необходима прежде всего для методики преподавания чтения и письма в школе. В ней нуждаются дикторы радио и телевидения, лекторы, актеры. Затранскрибированный текст даст возможность создать частотный словарь звуковых последовательностей казахского языка, который будет служить необходимым материалом для различных теоретических и прикладных исследований.

Казахская письменность построена на русском алфавите с некоторыми дополнениями, отражающими специфику языка. Она является в основном фонематической, т.е. каждая фонема обозначается в ней отдельной буквой. Однако имеются исключения. В речи некоторые фонемы подвергаются как комбинаторным, так и позиционным изменениям, в таких случаях буквы не соответствуют звукам.

Следовательно, для транскрипции казахского текста необходимо руководствоваться орфографическими правилами из 23 пунктов, которые подробно излагаются в докладе.

Работа является первым опытом автоматического получения транскрипционной фонетической записи казахских текстов на основе алгоритмов, реализованных на ЭВМ. Даются подробные фонетические правила реализации фонем. Описывается алгоритм преобразования письменных текстов в фонетическую запись.

Основные моменты преобразования казахских текстов в фонетическую запись заключаются в следующем:

1. Вручную отмечаются ритмические группы в предложениях. Далее идут преобразования, которые осуществляются при помощи ЭВМ.

2. Как показал анализ, дифтонги, трифтонги, обозначенные одной буквой вполне поддаются автоматическому делению на составляющие их звуковые компоненты.

3. Затем транскрибируются фонетические изменения, наблюдаемые внутри отдельного слова независимо от ритмического членения казахской речи.

4. Транскрибируются фонетические изменения, наблюдаемые внутри ритмической группы.

5. В последнюю очередь устанавливаются сингармонические (твердые, мягкие, губные, не губные) варианты фонем.

В результате выполнения программой 2-го пункта мы будем иметь такой текст, где каждому звуку соответствует отдельный символ, который еще не указывает на тот или другой вариант фонем.

Алгоритмизация задачи, связанной с 3-им пунктом, дают возможность внести в текст комбинаторные изменения фонем в отдельно взятых словах. Деление на ритмические группы помогает выявить фонетические изменения на стыках слов и сделать соответствующие корректировки в раннеполученный нами текст. В данном случае осуществляется транскрибирование позиционно-комбинаторной модификации фонем. Когда установлены все позиционно-комбинаторные изменения фонем, необходимо выявить сингармонически твердые и мягкие варианты фонем. Это осуществляется на основе свойственного для тюркских языков закона сингармонизма. Последним этапом автоматического получения фонетической транскрипции является внесение в текст корректировки касающейся сингармонически губных и негубных вариантов фонем, также специфичной для большинства тюркских языков, в том числе и для казахского языка. В результате выполнения всех этапов на выходе будем иметь окончательно протранскрибированный фонетический текст.

СОСТОЯНИЕ И ПЕРСПЕКТИВЫ ИССЛЕДОВАНИЯ ТЮРКСКИХ ЯЗЫКОВ С ПРИМЕНЕНИЕМ СТАТИСТИКО-ИНЖЕРНЫХ МЕТОДОВ

Отд. лит-ы и языка АН СССР. Отд. истор-и АН СССР.
Сов. комитет тюрк-в при ОЛЯ АН СССР АН Узб. ССР.
Всесоюз-я тюрк-кая конф-я: тезисы докладов и сооб-ий .
Ташкент, 10—12 сент. 1980 г. Изд. «Фан» Узб. ССР.
– С. 119-121. (Соав.: К.Б.Бектаев, К .Молдабеков и др.)

По словам известного тюрколога Н.А.Баскакова, до настоящего времени подавляющее большинство тюркологов-лингвистов имели предметом своего исследования описание строя конкретных тюркских языков, их лексической, грамматической и фонолого-фонетической структуры. В данное время в тюркологии настал этап перехода от описания строя конкретных тюркских языков в различных уровнях к описаниям типов языков, системы языков с помощью статистико-инженерных приемов.

Основная задача статистико-инженерного исследования текста заключается в выявлении статических характеристик относительно разных стилей и языков, в сопоставлении этих характеристик, в определении сферы действия этих характеристик в тюркских и других языках. Экспериментальной основой этих исследований являются частотные словари, составленные для разных языков и стилей по единой методике, и разработанные для единого вероятностно-статистического аппарата сравнительно-сопоставительного изучения разноструктурных языков. С этой целью составлены частотные словари разного вида; газетной лексики (казахского, узбекского, киргизского и каракалпакского языков), научно-технических текстов (узбекских, казахских), повестей и очерков (казахских, узбекских), детской литературы (узбекской, казахской) романов известных

писателей (казахских, узбекских) и фонемно-фонологические частотные списки (узбекского, казахского, каракалпакского, ногайского языков), общим объемом около двух миллионов словоупотреблений.

На базе анализа этих словарей получены известные результаты по статистико-информационному моделированию, которые успешно внедрются в типологические исследования, а также в исследования по искусственному интеллекту и машинному переводу и способствуют статической оптимизации. В частности, выявлены общеязыковые универсалии, каковыми являются статическая покрываемость (от 62 процентов до 89 процентов) и информационная покрываемость (от 55 процентов до 78 процентов) наиболее частых; слов, распределение средней длины слов в языках мира и т.п.

Моделирование статистической структуры тюркских текстов в виде частотных словарей различных стилей и подъязыков и толкование вероятностных характеристик конкретных слов разных лексико-грамматических классов и их сочетаний даёт возможность: а) выявить пласт общеупотребительной лексики; б) обнаружить слова, которые выступают в качестве дифференциальных, диагностирующих признаков каждого функционального стиля; а) использовать статистические критерии для выявления квантитативных типологических и стилистических характеристик.

ИСПОЛЬЗОВАНИЕ ЭВМ В АВТОРСКОЙ ЛЕКСИКОГРАФИИ

Мин-во высш. и сред. спец-го обр-ния СССР и БССР
Минский Гос. Пед. Институт иностр-х языков.
Тез-ы докладов Всесоз. конф-ии: «Переработка
текста метод-ми инж-ной лингв-ки». Минск, 1-2 февраля
1982 г. – С. 135-137. (Соав.: К.Б.Бектаев и др.)

Одной из главных задач языкознания является всемерное расширение одноязычных и многоязычных лексикографических работ, в том числе оставление толковых словарей по отдельным произведениям и по полному собранию сочинений известных писателей и мыслителей. Ср. «Словарь языка А.С.Пушкина» (1956, 1957, 1959, 1961 гг.), «Словарь языка Абая» (1968), «Словарь языка В.И.Ленина», составляемый Институтом русского языка АН СССР. Обязательным условием создания этих контестуальных словарей является отбор слов с его контекстным окружением (словосочетание, предложение, ряд предложений). До настоящего времени такие словари создавались путем ручного составления карточек, тому во многих языковых институтах накопились миллионы, а зачастую десятки миллионов карточек. Например, для составления «Словаря языка Абая» по текстам объемом около 50 тысяч словоупотреблений пришлось составить 50 тысяч карточек, сортировать их по алфавиту, после чего формулировать толкование слов. Для выполнения этой работы понадобилось около 5 лет труда десяти работников. Следовательно, для составления такого словаря по полному собранию

сочинений в 20-ти томах (текст объемом около 2,5 миллионов словоупотреблений) выдающегося казахского писателя и драматурга М.Ауэзова этому же коллективу добилось бы минимум 50 лет.

Около 70% лексикографических работ являются нетворческими, механическими процедурами, которые вполне можно поручить ЭВМ. В связи с этим коллектив группы «Статистико-лингвистические исследования и автоматизация» ИЯ АН Каз. ССР и Чимкентский пединститут разработали алгоритм и программы составления АЧРСС с помощью ЕС-ЭВМ. Как показывает опыт, составитель толкового словаря, имея при себе АЧРСС, легко разыскивает и находит по порядковому номеру части страницы и страницы тома слова с их контекстами. Подготовка исходных данных осуществляется на магнитном накопителе, при перфорации используется дисплей.

В Дагестанском госуниверситете создается аварско-русский ЧС языка произведений народного поэта Дагестана Расула Гамзатова. Словарь составляется по шеститомному изданию избранных произведений автора с подбором соответствующих переводов на русский язык. Объем лексического запаса в указанном издании составляет около 1,5 млн. словоупотреблений.

Весь объем информации наносится на МЛ с использованием устройства ЕС-9002.

Для составления ЧС с последующим выводом его на АЦПУ используется программа простого перебора, расположения в алфавитном порядке. Для этого информация считывается поблочно, а затем происходит преобразование в дословные элементы, с которыми и ведется работа.

Для получения картотечной базы по произведениям автора (карточки-цитаты) информация считывается с МЛ поблочно, затем преобразуется в побуквенные элементы: выделяются предложения, которые сравниваются с ключевыми словами. Присутствие в предложении одного из ключевых слов говорит о необходимости вывести данное предложение на внешнее устройство. В качестве него используется МД, на которых создается индексно-последовательный файл, где ключами являются слова, которым составляется картотека. Для нахождения иформации необходимо задать только слово-ключ, и по нему прямым доступом находится информация.

Предполагается получить с помощью ЭВМ АЧС и картотечную базу по каждому произведению автора, распределительные словари (по произведениям, томом), и, наконец, сводный алфавитно-частотный аварско-русский словарь.

Несколько в ином плане используется ЭВМ в Алма-Атинском педагогическом институте иностранных языков (АПИИЯ). Целью такого исследования являлось выявление специфических лингвистических элементов, характеризующих как речь автора, так и речь отдельных литературных персонажей определенного художественного произведения. Материалом исследования послужил роман немецкого писателя Д.Нолля «Приключения Вернера Кольта» длиной около 150 тыс. словоупотреблений. Текст был обработан в Лаборатории статистики речи АПИИЯ.

Для определения степени статистической близости текстов одной выборки весь роман был условно разделен на две части: авторская речь и речь отдельных литературных персонажей. Речь главного героя произведения анализировалась в двух направлениях: собственно речь героя и его мысли (внутренний монолог).

В области грамматики были определены соотношения времен и их грамматических средств выражения, включая пассив и условное наклонение. Обращено внимание на употребление прилагательных в форме наречий. В особую группу выделены так называемые реалии и имена собственные.

СТАТИСТИЧЕСКИЙ АНАЛИЗ ГРАММАТИЧЕСКИХ ОФОРМИТЕЛЕЙ НА ЭВМ

АН СССР Отд-ие лит-ы и языка. Советский комитет тюркологов. Отделение истории АН Киргизской ССР. Тезисы докл-в и сооб-ий V Всесоюзной тюрколог-кой конф-ции: «Тюркология-88». (7–9 сентября 1988 г.). Фрунзе: «Илим»-1988.

Количественные критерии обеспечивают углубленное изучение потока речи как опознанную действительность. Важное значение в этом приобретают составляемые с помощью ЭВМ различные частотные словари, на базе которых выявляются определенные языковые закономерности.

Исходя из этого, нами проводятся статистическое исследование морфемного состава различных словоформ казахского языка, выявляется ряд структурных особенностей формирования грамматических классов слов на уровне морфологии агглютинирующего языка. Статистико-морфологическому анализу подвергаются пути грамматикализации имен существительных, глагола, прилагательных и местоимений из казахских художественных и публицистических текстов.

Главный упор делается на обратно-алфавитные частотные словари как рациональные пособия для зримого учета и дальнейшего вычленения словообразовательных и формообразовательных моделей и типов, отчетливо представляющих сочетаемость того или иного аффикса с различными морфемными образованиями, наглядно иллюстрирующих функции определенного аффикса в тех или иных грамматических наращениях.

В казахских текстах, по сравнению с флективными языками, налицо ярко выраженная формообразовательная способность агглютинирующего строя, которая обнаруживается в покрываемости текста. Так, если в индоевропейских языках 100-150 высокочастотных словоформ покрывают 50% текста, то в казахском языке указанный процент текстового массива покрывается более чем 800 словоформами. Падежные, притяжательные и личные окончания, а также окончания множественности, прибавляются к слову в строго определенном порядке. Количественно наращивается на более

трех оформителей. Но все четыре типа окончаний одновременно не употребляются в составе одной словоформы, что связано с внутренними морфологическими и фонетическими закономерностями языка.

Выявлено, что в тексте романа М.Ауэзова «Абай жолы», состоящего из 466 тыс. словоупотреблений, 64% производных глаголов по своему генезису – отглагольные, из числа которых 91% образован при помощи залоговых аффиксов. Причем одна глагольная основа может наращиваться одним, двумя, тремя и более залоговыми аффиксами, что является специфической особенностью.

Используя обратно-алфавитный словарь, нами установлено, что имя прилагательное в количественном отношении и в художественном, и публицистическом текстах занимает третье место после имен существительных и глагола. Особый интерес представляет употребление писателем степеней сравнения прилагательных, которые образуются при помощи собственных аффиксальных элементов. Отмечено, что преобладание основной или производной форм слова в тексте зависит от того, к какой части речи она относится, а также от того, к какому жанру принадлежит текст. Например, такие менее подверженные грамматическим изменениям части речи, как наречия, междометия, а также служебные и подражательные слова, сохранили в текстах в основном свои исходные формы.

Статистический анализ грамматических оформителей показал, что агглютинирующий казахский язык имеет более развитую и многообразную систему грамматикализации в отличие от языков с флективным и аналитическим строем, выяснилось, что для публицистического и художественного стилей присущи специфичные фонетические варианты грамматических форм, характеризующихся, своими частотами употребления.

Полученные статистические данные могут быть использованы при составлении практических курсов по морфологии казахского языка. Они необходимы для разработки алгоритмов морфемного членения слов в целях создания эффективной программы автоматического выделения морфологических типов методами статистико-комбинаторного анализа.

О СОЗДАНИИ МАШИННОГО ФОНДА КАЗАХСКОГО ЯЗЫКА

АН груз. ССР. ЦНИ по общественным наукам.
Матер. раб-го совещ-я. «Машинные фонды
языков народов СССР». Тбилиси, 13-22 ноября
1988 г. – С. 16-17. (соав. – С.А.Ускомбаев).

1. Применяемые математические и кибернетические методы в индоевропейском языкознании оказали свое влияние и на иследования в области тюркологии. При функционировании в тюркских языках многие их элементы мало подвержены фонетическим и морфологическим изменениям, сохраняя при этом «прозрачность» своей структуры. Это дает возможность

точно и недвусмысленно фиксировать те или иные моменты языка при математическом описании и автоматизации тюркологических исследований.

2. В решении Второй Всесоюзной конференции по созданию Машинного фонда русского языка, учитывая опыт и объем работ, выполняемых казахстанскими исследователями по машинной обработке текстов тюркских языков, предложено назначить организатором-координатором по созданию машинных фондов тюркских языков Институт языкознания АН Каз.ССР.

3. В Институте языкознания АН Каз.ССР уже ряд лет осуществляются работы по созданию иллюстративно-текстового фонда казахского языка. В данное время проводятся работы, связанные с первым этапом, т.е. на машинные носители переносятся тексты 20-ти томного собрания сочинений М.О.Ауэзова. Параллельно составляются на ЭВМ алфавитно-частотные словари-словоуказатели по каждому тому. В словаре словоформы, кроме частоты употребления, снабжаются номерами страниц и строк, где они встретились в книге. По желанию лингвиста-исследователя можно увидеть любой контекст с интересующей словоформой на экране дисплея или получить распечатанную на бумаге карточку – цитату с опорным словом или словоформой. Сейчас такие словари-словоуказатели получены на ЭВМ по первым 10-ти томам. К 1990 году запланировано закончить перенесение на магнитную ленту всего корпуса текстов полного собрания сочинении (объемом около 3 млн. словоупотребление) и получение на ЭВМ словарей-словоуказателей как по отдельным томам, так и в целом по полному собранию сочинений.

Отмечен и второй этап создания машинного фонда казахского языка. Это – обработка на ЭВМ текста 10-ти томного толкового словари казахского языка с предварительной разметкой реестровых слов по лексическим значениям, морфологической структуре и синтаксическим функциям.

4. В перспективе думается включить в машинный Фонд казахского языка Генеральный словник казахских словарей, который можеть быть создан как семейство уже имеющегося словаря по полному собранию сочинений произведений М.Ауэзова, 10-ти томного толкового словаря, казахского языка, грамматического словаря, частотных словарей по разным жанрам и других «словникоподобных» материалов.

Одним из компонентов Машинного фонда казахского языка будет терминологический фонд, который имеет уже машинную форму в отделе терминологии и теории перевода Института языкознания АН Каз.ССР. Другим компонентом Машинного фонда казахского языка в перспективе должен стать Академический словарно-грамматический фонд, который будет состоять из нескольких подфондов. Главным из них станет свод академических словарей, выполняющий не только справочные функции, но и являющийся источником нового систематического изучения лексики. Академический словарно-грамматический фонд будет расширяться за счет историко-этимологического, диалектологического, топонимического и других подфондов.

5. Машинный фонд любого языка должен иметь достаточное программное обеспечение. Это – программные комплексы, реализующие автоматический морфологический, синтаксический и семантический анализ текстов, т.е. так называемые языковые процессоры, разработка которых планируется в ближайшем будущем.

6. Для успешного осуществления проблемной задачи – создания машинных фондов тюркских языков, во-первых, необходимо концентрирование усилий многих научных коллективов тюркоязычных республик, использующие методы автоматизации лингвистических работ. Во-вторых, создателям машинных фондов тюркских языков необходимо перенять опыт научных коллективов, которые создают Машинные фонды русского, украинского, эстонского и других, языков.

О ПРОБЛЕМЕ СОЗДАНИЯ МАШИННОГО ФОНДА ТЮРКСКИХ ЯЗЫКОВ

АН Эстонской. ССР Институт языка и литературы.
Матер-ы рабочего совещания: «Машинные фонды языков народов СССР». (Таллинн, 19-22 декабря 1988) – С. 4-5.

На XIV Пленуме Комитета советских тюркологов (Москва, 1988) отмечены необходимость создания общетюркского машинного фонда, пути проектирования и разработки, лингвистические концепций – научного и многоцелевого МФТЯ. Центром МФТЯ рекомендован Институт языкознания АН Каз.ССР. Для изучения процесса порождения грамматических форм тюркского слова на начальном этапе целесообразнее создание МФТЯ, включающий полный набор структурно-фонетических разновидностей односложного слова. В дальнейшем планируется создание большого многоцелевого МФТЯ, моделирующий как общетюркскую систему, так и систему каждого конкретного языка. МФТЯ способствует накоплению и упорядочению в сравнительно-сопоставительном плане лексикографические, грамматические, историко-этимологические сведения о языках.

ПОДГОТОВКА ТЮРКОЯЗЫЧНОГО ТЕКСТА ДЛЯ ПОЛУЧЕНИЯ СЛОВОУКАЗАТЕЛЯ ПРОИЗВЕДЕНИЙ ИНДИВИДУАЛЬНОГО ПИСАТЕЛЯ НА ЭВМ

АН СССР АН Азербайджанской ССР
«Советская тюркология». № 1,
Баку – 1988. 115-122-бб.

Научно-технический прогресс, совершенствование электронно-вычислительной техники, внедрение точных методов исследования в

различные области знания оказывают интенсивное влияние и на развитие лингвистической науки, в том числе тюркологии.

В данной статье анализируется некоторый опыт применения электронно-вычислительной машины (ЭВМ) в лингвистических изысканиях, осуществляемых в течение ряда лет в Институте языкознания АН КазССР. Как показывает практика, с помощью ЭВМ можно значительно облегчить наиболее трудоемкие процессы выполнения лексикографических работ, в частности заменить быстродействующими и удобными магнитными лентами, перфокартами и т. д. нынешние громоздкие картотеки, создаваемые вручную. В результате во много раз упрощаются поиск и накопление необходимой языковой информации, становится возможным ее долгосрочное хранение.

В настоящее время полностью автоматизирован процесс построения различных видов частотных словарей, таких, как алфавитно-частотные, частотно-алфавитные и обратно-алфавитные словари лингвистических единиц текста. Подобные словари можно получать с помощью ЭВМ по текстам и тюркских языков. При этом появляется возможность сопоставить большой статистический материал, накопленный по индоевропейским языкам флективно-аналитического строя, со свежими статистическими данными языков агглютинативного строя. Опыт показал, что строй тюркских языков поддается эффективной обработке на ЭВМ, поскольку морфемная структура тюркских основ по своей фонетической и грамматической природе является довольно устойчивой и прозрачной. Как считают Р. Г. Пиотровский и другие ученые, ассимиляция, редукция и прочие явления, входящие в понятие «фонетических законов» в рамках индоевропейских языков, чужды тюркским языкам, а это позволяет точно фиксировать те или иные моменты языка при математическом описании.

Современное развитие казахского языкознания предполагает учет не только фонетических, грамматических классов слов, но и всего словарного фонда языка как в литературном, так и в диалектальном аспектах. Ныне Институт языкознания АН КазССР располагает обширным картотечным фондом, который ежегодно пополняется за счет словоформ, почерпнутых из новых изданий. Благодаря изучению такого материала созданы и создаются орфографический, терминологический, двуязычные, а также 10-томный «Толковый словарь казахского языка». Однако трудоемкость ручной выборки материалов картотечного фонда создает немало неудобств при выполнении плановых заданий как в теоретико-исследовательском, так и в организационном, практическом плане.

В Институте языкознания АН КазССР ряд лет функционирует группа статистико-лингвистических исследований и автоматизации (ныне группа автоматизации лингвистических работ), которая занимается внедрением автоматизации в выполнение отдельных трудоемких процессов в области лексикографии. Группа использует статистические методы исследования как необходимое условие для разрешения многих вопросов не только структурно-прикладного, но и традиционного языкознания.

При помощи ЭВМ сотрудники группы подготовили и опубликовали частотные словари: «Куманша-қазақша жиілік сөздік» («Куманско-казахский частотный словарь») и «М. Әуезовтің «Абай жолы» романының жиілік сөздігі» («Частотный словарь романа М. Ауэзова «Путь Абая») [1. С. 277; 2. С. 336]. Материалы этих словарей могут помочь языковедам при исследованиях количественных характеристик лингвистических фактов.

Практика работы с частотными словарями показывает, что алфавитно-частотные словари слов или словоформ дают больше лингвистической информации, если слова словника будут снабжены «адресами», т. е. указанием страниц, где они встречаются. Это упрощает задачу исследователя при ознакомлении с интересующим его словом или словоформой из контекста. В этой связи возникла необходимость составить на ЭВМ словарь-словоуказатель с выдачей для каждой словоформы нумерации страниц и строк по их встречаемости и частоте в исследуемом тексте.

В настоящее время в Институте языкознания АН КазССР началась подготовка к составлению частотного словаря языка произведений М.О.Ауэзова на базе полного собрания его сочинений в 20-ти томах.

Полный состав языка произведений М. Ауэзова включает в себя около 3 млн словоупотреблений. При таком большом объеме текста невозможно зарегистрировать весь лексический материал вручную. К объективным трудностям относятся сложность материала, содержащего большое число (более 50% текста) редких слов по отношению к частоте их употребления, а также множество однотипных, часто употребляемых слов, среди которых есть и лексические аномалии, сочетания и т. п. Как правило, наличие большого числа определенных элементов слов с высокой частотой употребления, например, предлогов, союзов, местоимений, вспомогательных глаголов и т. п., создает известные трудности при обработке. С другой стороны, полная регистрация указанных лингвистических единиц приводит к обилию однотипного материала, который заполняет все емкости картотеки и затрудняет в дальнейшем работу составителей. В некоторых случаях выборщик поступает и так: если тот или иной контекст по какой-либо причине не удовлетворяет его, то он пропускает этот синтаксический отрезок, хотя, может быть, именно в нем употреблено слово, встречающееся во всем произведении всего один раз. Это ведет к немалым потерям слов. Таким образом, полная обработка текста произведений на ЭВМ дает возможность зарегистрировать словарный состав источников с охватом всех лексических единиц текста, с последующим отбором необходимых для словарной работы языковых величин. Все это способствует успешному решению вопросов, связанных с полнотой словника словаря, нормативностью фиксируемых лексических единиц, их стилистической характеристикой в словаре, позволяет отделить узуальное от индивидуально-авторского и т. п. Практика работы с автоматическими устройствами показывает, что выполнение такого трудоемкого процесса, как полная обработка текста, возможна только при использовании ЭВМ.

Основная трудность обработки лингвистических данных с помощью электронной техники заключается в подготовке необходимой информации для ввода в ЭВМ, т. е. в перфорации текста. Пока этот процесс еще не автоматизирован, отсутствуют и читающие устройства, и потому перфорирование в данное время ведется вручную, что, в свою очередь, влечет за собой и неизбежные ошибки, требующие неоднократной проверки подготовительного периода операций.

Не исключено, что из-за недостаточности формальных признаков нужной лингвистической единицы становится невозможным автоматическое выделение ее из среды текстового массива. В подобных случаях возникает необходимость в предварительном снабжении таковой условными индексами типа помет. Например, для выделения синонимических, омонимических и антонимических рядов из состава художественного произведения или получения списков лексических групп типа архаизмов, неологизмов, модернизмов и других нам приходилось прибегать к индексации текстового материала.

При подготовке текстов 20-томного собрания сочинений М. Ауэзова для ввода в ЭВМ нами приняты условные обозначения собственных имен через буквенные индексы, а именно: А – имена и фамилии; Ж – клички животных; Г – географические названия; Б – названия печатных изданий; М – исторические события, культурно-бытовые названия; Н – астрономические названия; Р – названия племен и родов; С – аббревиатуры из заглавных букв. Индексы ставились в начале слов через дефис в расчете на то, что собственные имена в алфавитно-частотном словаре будут выделены особо. Были помечены также встречающиеся в тексте фразеологизмы, пословицы и поговорки, которые в дальнейшем мы намерены выделить в виде отдельных списков программным путем.

Снабжение текста индексами желательно проводить при решении максимального количества лингвистических задач, так как перфорирование подготовленного материала, как уже отмечалось, является не менее трудоемким делом.

Таблица 1

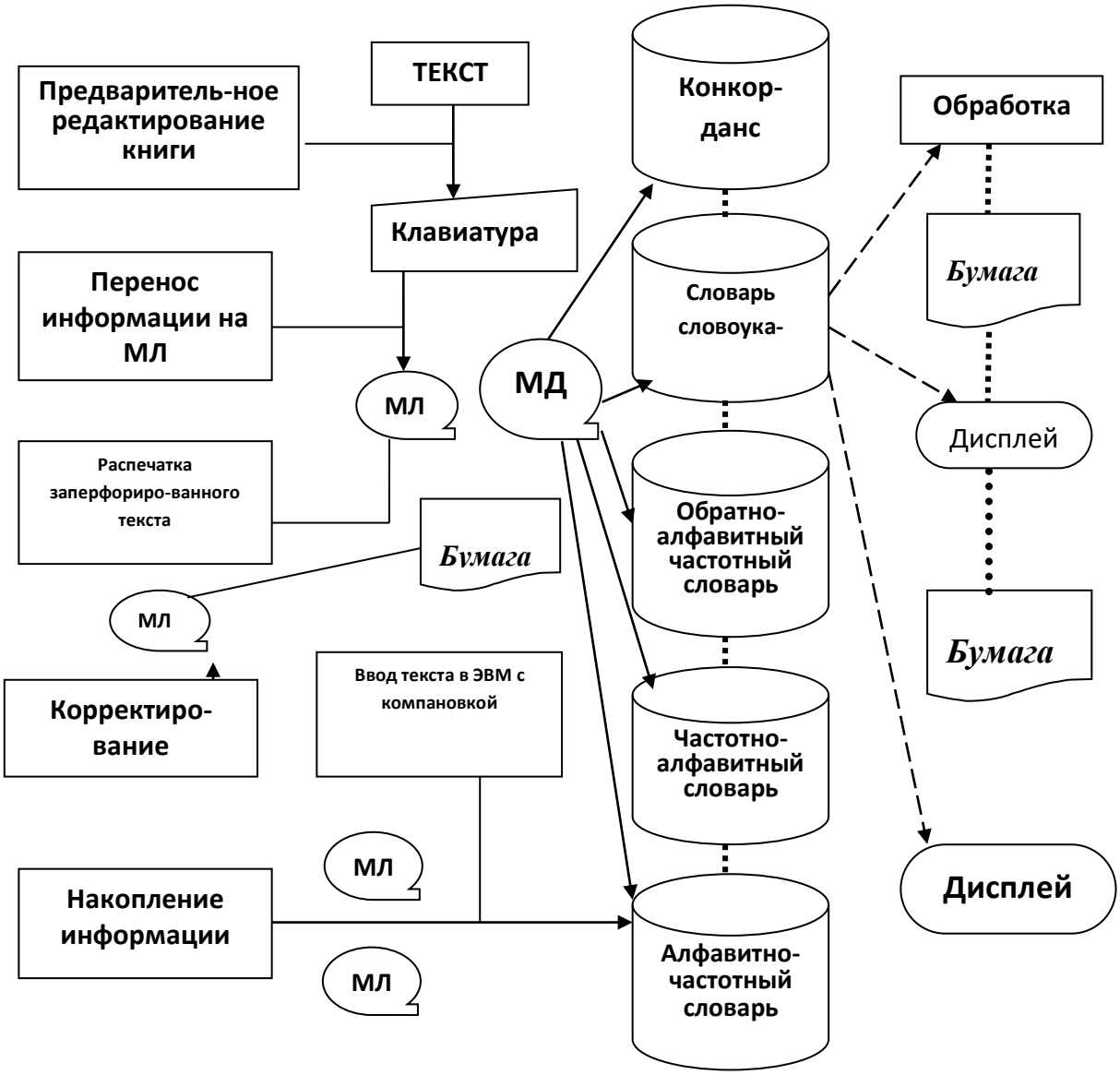
Единная транскрипция специфических букв тюркских языков

Азебайджанский	Алтайский	Башкирский	Гагаузский	Казахский	Каракалпакский	Карачаево-Балкарский	Киргизский	Татарский	Тувинский	Туркменский	Узбекский	Уйгурский	Хакасский	Чувашский	Якутский	Транскрипция
			Ä											Ă		@
Ғ		Ғ		Ғ	Ғ						Ғ	Ғ	Ғ		ҕ	F
								Җ		җ		җ				G
Қ				қ							қ					Q

	ң	Ң		ң	ң		ң	Ң	ң	ң		ң	ң			N
ө	ö	ө	ö	ө	ө		ө	ө	ө	ө		ө	ö		ө	¤
		Ç												ç		S
				Ұ										Y		U
ә		ә		ә	ә					ә		ә				Э
h		H		h	x			h		h	x	h			h	&
				I									I			I
Y		Y		Y	Y		Y		Y	Y		Y			Y	Y
J	J															J

Как известно, аппараты, кодирующие текстовую информацию, вводные и выводные устройства современных ЭВМ основаны на латинской и русской графике. Следовательно, при кодировании тюркоязычных письменных текстов необходимо учитывать и графемы с диакритическими знаками, принятыми в соответствии с фонематическими особенностями того или иного языка.

Схема организации обработки текста на ЭВМ



Для удобства кодировки таких текстов и универсальности машинных программ, предназначенных для обработки лингвистического материала изучаемых языков, мы применили единые обозначения идентичных фонем различных тюркских языков. Наша транскрипция преследует узкую цель — ввод текстовой информации тюркских языков в память ЭВМ (табл. 1).Как видно из таблицы, принятые нами буквозаместители для обозначения специфических фонем позволяют беспрепятственно пользоваться кодировочными аппаратами, основанными на русской и латинской графике.

Принцип транскрибирования специфических букв тюркских языков позволяет вводить в ЭВМ и фонетический текст, что необходимо, для исследования фонетической структуры языка с последующим определением фонематического состава, частотности индивидуальных фонем, а также рядов их позиционных вариантов и т. д.

На наш взгляд, имеющиеся алгоритмические программы нуждаются в частичном пересмотре в связи с особенностями тюркских языков, предполагающих наличие специфических букв, а также имеющих свой порядок расположения букв по алфавиту. Учитывая все это, мы составили для унификации работы машинный алфавит, общий для целого ряда тюркских языков, таких, как казахский, узбекский, туркменский, азербайджанский, киргизский, каракалпакский и др. [3, С. 250–262].

В целом организацию обработки текстовой информации на ЭВМ и получение различных видов словарей можно представить в виде схемы, показанной на рисунке.

Предварительная подготовка обрабатываемого материала в наших опытах состояла из следующих этапов: а) исправление орфографических ошибок в тексте, допущенных по вине издательств; б) индексирование слов, означающих имена собственные; в) выделение соответствующими условными знаками фразеологических сочетаний, двух-, трех- и многочленных единиц, пословиц и поговорок; г) соединение специальным знаком компонентов сочетаний, не употребляющихся раздельно; д) разделение текста каждой страницы книги на условные строки и их нумерация (деление текста страницы на смысловые абзацы).

Подготовленный таким образом текст переносится оператором на магнитную ленту (МЛ) с помощью специального устройства подготовки данных (УПДЛ) ЕС-9004. Запись на МЛ ведется по следующей программе (см. табл. 2).

Данная программа позволяет в одном блоке МЛ записывать 70 символов текста и 10 резервных знаков «звездочка» (*). Корректировка записи, предполагающая ее удлинение, осуществляется за счет указанных резервных позиций. При этом перед началом работы в позицию 71–80 буферной памяти устройства необходимо занести символ «звездочка» (*). Если же выделенных для резерва 10-ти позиций недостаточно, то допускается переперфорирование соседних записей. В этом случае резервные

позиции соответственно увеличиваются. Остальные моменты работы можно вести согласно инструкции по эксплуатации устройства ЕС-9004.

Таблица 2

Реквизит	Программный код	Номер позиции	*Количество* позиций
Текст	КИР	1-70	70
*****-резервные	ДБЛ	71	1
позиции	КИР	72-80	9

Номера страниц и строк перфорируются в начале записи со 2-й позиции перед каждым абзацем в виде (24-1), (24-2), (27-3) и т.д., заключенных между пробелами так же, как и любое слово текста. Круглые скобки не должны использоваться в других случаях, поэтому они заменяются иными видами скобок. В скобке перед дефисом стоит число, означающее номер страницы, а после дефиса – номер строки.

Скорость перфорации казахских текстов в среднем за один рабочий день пока равна 2,5 тыс. словоупотреблений, или 8 страницам книги. Относительно невысокая продуктивность оператора, видимо, объясняется тем, что в казахских текстах часто встречаются специфические буквы, которые передаются через буквы латинского алфавита и другие служебные знаки. Следовательно, в процессе перфорации для записи одной казахской буквы на МЛ оператору приходится нажимать две клавиши. Для увеличения скорости операции требуется инженерная переделка клавиатуры устройства согласно казахской графике письма.

В схеме на рисунке указаны этапы организации обработки текста на ЭВМ. Следующим этапом после записи определенной порции текста на МЛ является распечатка ее на бумаге. Она необходима для обнаружения ошибок, допущенных во время перфорации.

Исправление опечаток в тексте корректируется тем же устройством ЕС-9004. Этот процесс значительно облегчен благодаря использованию экрана дисплея в устройстве для чтения блока данных на МЛ. Исправленная окончательно порция текста дозаписывается на накопительную МЛ. Для повышения надежности накопительная МЛ должна храниться в 2-х или 3-х экземплярах. Данный этап работы повторяется до тех пор, пока на накопительной МЛ не окажется весь необходимый объем текстовой информации.

Комплекс программ для получения алфавитно-частотного словаря-словоуказателя на ЭВМ написан на алгоритмическом языке ПЛ/1 для ЕС ЭВМ. Для получения словаря-словоуказателя необходима в основном работа 3-х программ в следующей последовательности:

1. Ввод текста с МЛ на магнитный диск (МД);
2. Сортировка словаря по алфавиту, номерам страниц и строк;
3. Печать на АЦПУ словаря-словоуказателя.

Программа вывода на печать словаря считывает с магнитного диска записи и печатает на устройстве печати (АЦПУ) словоуказатель с указанием

порядкового номера, частоты употребления, словоформы, номеров страниц и строки.

Для каждой словоформы по данной программе в памяти ЭВМ зарезервировано место на 500 номеров страниц. Так, если словоформа встречается более чем 500 раз, то печатаются только первые 500 номеров страниц. Такое ограничение вытекает из предположения, что данная словоформа зафиксирована почти на всех страницах книги, объем которой не более 500 страниц, что исключает необходимость их указания. Образец словаря-словоуказателя приведен в табл. 3.

Ресурсы ЭВМ ЕС-1045, необходимые для обработки одной книги объемом 350 страниц (около 120 тыс. словоупотреблений), таковы: оперативная память – 160 кбайт; память на магнитной ленте – 1 катушка МЛ (350 м); магнитные диски – 10 мбайт; машинное время – 2 ч.

Таблица 3

Собрание сочинений М. О. Ауэзова. Т. I			
№ п.п.	**Абсолютная частота**	**Словоформа**	**Номера страниц и строк**
1	9	А	(87-4) (144-6) (226-1) (226-4) (226-7) (262-1)(315-6) (345-3) (346-1)
2	1	А-АБАЙДЫN	(177-4)
551	3	АҒА	(210-1) (227-1) (248-6)
617	68	АДАМ	(68-2) (72-3) (73-7) (74-1) (78-2) (78-8) (79-1) (108-1) (128-7) (133-2) (148-5) (149-2) – – – (376-8) (380-1)
2813	5	БАQЫТ	(130-7) (166-3) (169-2) (209-5)(252-2)
5008	1	ДUҒА	(252-2)
5360	5	ЕNБЕГІ	(289-4) (302-5) (334-1) (336-8) (348-5)
7313	14	ЖUМСАQ	(74-2) (91-5) (123-6) (145-3) (159-2) (166-3) (168-4) (170-2) (179-7) (236-1) (241-4) (244-7) (348-6) (348-6)
8783	1	КОЗІNНІN	(78-1)
9915	4	QАЗАQЫ	(102-7) (306-) (318-1) (332-2)
14348	6	PUQCAT	(140-3) (254-4) (295-2) (278-5) (299-2) (333-8)
15408	1	СОНІП-ЖАНУ	(142-1)
19971	3	ЯПЫР-АУ	(199-3) (333-5) (398-3)

Таким образом, на ЭВМ будут получены словоуказатели как по отдельным томам, так и в целом по полному собранию сочинений. Предполагается получение и объединенного распределительного (по частоте) алфавитно-частотного словаря, что позволяет выявить распределение частот слов в отдельных томах сочинений М. О. Ауэзова. Кроме вышеназванных словарей, на ЭВМ будут получены частотные и обратно-алфавитные словари, а также списки фразеологизмов, включающих пословицы и поговорки, встретившиеся в произведениях М. О. Ауэзова.

В данное время сотрудники группы автоматизации лингвистических работ Института языкознания АН ҚазССР подготовили тексты восьми томов, и на ЭВМ по этим томам получены частотные словари-словоуказатели.

Хранение на магнитной ленте текста полного собрания сочинений М.О.Ауэзова позволит благодаря ЭВМ вывести на печать или показать на экране дисплея любой абзац текста с интересующим исследователя словом, а словарь-словоуказатель укажет на страницу и строку данного слова в книге.

Нам думается, что полученный таким образом на ЭВМ справочный материал по всем произведениям М.О.Ауэзова послужит ценным источником не только для исследователей языка писателя и составителей толкового словаря языка М.О.Ауэзова, но и для изучения различных проблем современного казахского языкознания.

Литература

1. *Құрышжанов А. Қ., Джубанов А. Х., Белботаев А. Б.* Куманша-қазақша жиілік сөздік. Алматы: Фан, 1978.

2. *Бектаев К. Б., Мырзабекоа С., Белботаев А. Б.* М.Әуезовтің «Абай жолы» романының жиілік сөздігі. Алматы, Фан, 1979.

3. *Бектаев К.Б., Джубанов А. Х.* Индексация и кодирование текстовой информации для ввода в ЭВМ//Статистика казахского текста. Алма-Ата: Наука, 1973.

СОЗДАНИЕ СЛОВАРНОЙ БАЗЫ ДАННЫХ ОБЩЕ– И МЕЖТЮРКСКИХ МОНОСИЛЛАБОВ: К ПРОБЛЕМЕ РАЗРАБОТКИ МАШИННОГО ФОНДА ТЮРКСКИХ ЯЗЫКОВ

АН СССР Институт русского языка.
Третья Всесоюзная конференция
по созданию машинного фонда
русского языка. Тезисы докладов,
Москва, 1989. – Ч. II. – С. 129-132.
(Соав.: Е.З. Кажибеков, С.А.Ускомбаев)

1. На современном этапе развития тюркологии все еще недостаточно разработаны такие разделы лексикологии, как создание полных словарей каждого из конкретных живых тюркских языков, а также мертвых полных словарей древне– и среднетюркских и др. памятников диалектной лексики.

Исходя из того, что односложные корневые слова являются лексической базой тюркских языков, выяснение исчерпывающего состава корневой лексики по конкретным тюркским языкам, определение специфики ее содержания и объединение их в одно целое представляет собой важную задачу исторической лексикологии.

Разнообразный по своему составу и структуре фонд исконной тюркской лексики, включающий двух-, трех - и многосложные единицы и

количественно превышающий однословные корни-основы в каждом из тюркских языков, является результатом деривационного развития моносиллабов.

Словарный репертуар современного казахского языка, выявленный на базе зондирования наличных словарей, составляет более 70 тыс. реестровых слов, из них около 3 тыс. (2704) единиц приходится на долю однословных корней и основ. Такие данные было бы желательно иметь и по другим тюрксим языкам.

Изучение односложных корней и основ путем сведения их в одну общую систему – один из возможных приемов изучения этой категорий лексем. Такой подход имеет ряд преимуществ перед разрозненными эмпирическими, фрагментарными исследованиями [1; 2; 3].

2. С 1990 г. по 2005 г. намечается проведение исследований по Комплексной научной программе «Проблемы реконструкции тюркского праязыка» (V-II тыс. до н.э.).

В процессе реализации Программы по отбору и систематизации материала будет рассматриваться проблема структурного соотношения пратюркских ядерных, общих и периферийных, диалектных черт в их движений и взаимодействии. Членение фактологического материала на семантические группы и их углубленное изучение будет осуществлено после завершения свода воедино первой лексикографической части всего корпуса пратюркского словарного фонда.

Предстоит решение задачи создания прасловарей на базе конкретных языков, их групп, объединение и учет всех показателей, снабженных этимологическим комментарием. Будет осуществляться праформно-гнездовой и генетический принципы распределения лексем, частеречный принцип и семантическая дифференциация материала.

3. В мае 1988 г. в Москве состоялся XIV Пленум Советского Комитета тюркологов. На повестке дня Пленума специально был поставлен проблемный вопрос о создании Машинного фонда тюркских языков (МФ ТЯ). Докладчики (В.Г.Гузев, А.М.Щербак, Р.Г.Пиотровский) отметили, что в МФ ТЯ должны быть собраны и упорядочены в сравнительно сопоставительном плане лексикографические, грамматические, статистико-стилистические, историко-этимологические сведения, а с другой стороны – заданы правила порождения и выдачи потребителю корректных лингвистических единиц конкретных тюркских языков.

Были предложены две стратегии создания МФ ТЯ (НМФ ТЯ), призванный обслуживать многоцелевой МФ ТЯ. Было рекомендовано функций головной организации по созданию МФ ТЯ возложить на Институт языкознания АН Каз.ССР [4].

Первоначальный научный МФ ТЯ (НМФ ТЯ), призванный обслуживать Комплексную научную Программу «Проблемы реконструкции тюркского праязыка», должен включать в себя исчерпывающую информацию о структурно-фонетических разновидностях тюркского односложного слова, наборы морфем и синтаксических схем,

грамматический тезаурус аффиксов и аналитических показателей. Таким образом, создание НМФ ТЯ предполагает создание лингвистического банка данных тюркских языков для изучения проблем фоне-морфо-лексико-семантического анализа обше- и межтюркских моносиллабов.

Одним из главных моментов при построении НМФ ТЯ является выяснение лингвистической концепции и связанных с ней перспектив реального использования Фонда, аппаратно-технической базы и стратегии программирования.

4. На втором этапе тюркологические научные центры страны должны заниматься созданием большого многоцелевого МФ. ТЯ или НМФ ТЯ, который должен строиться, моделируя как общетюркскую систему, так и систему каждого конкретного языка. НМФ ТЯ в основном обязан решать прикладные задачи, связанные с практическим использованием, или преподаванием тюркских языков.

Основными блоками НМФ ТЯ являются: 1) лексический блок, 2) грамматический блок, 3) фонологический и морфологическиq блоки.

4. Всю работу по созданию НМФ ТЯ необходимо осуществлять, придерживаясь нисходящего принципа, когда проектирование МФ и программирование ведется сверху вниз; это позволяет избежать опасности рассогласования модулей [4]. Считаем правильным начинать работу с обработки на ЭВМ материалов каждого конкретного языка. Только располагая достаточными данными, можно приступать к суммированию результатов с позиции общетюркского осмысления.

Работы, связанные с созданием МФ ТЯ, в перспективе будут вестись на уровне самостоятельной Всесоюзной программы, с точно определенным кругом исполнителей и мощной финансовой и технической базой.

5. Общетюркское осмысление полученных машинных результатов не может обойтись без сравнительного обследования фонетики тюркских языков. Для разработки графемно-фонематических программ следует иметь унифицированную орфографию и фонетическую транскрипцию тюркских языков, которые необходимы и на предыдущих этапах лексикографического накопления фактографического материала. Машинная обработка тюркской речи (текста) должна осуществляться с учетом сингармонизма – основной фонологической закономерности тюркских языков.

Следует заняться решением вопросов, связанных с автоматическим фонетическим транскрибированием различных тюркских текстов.

Высказанные нами соображения по поводу проектирования МФ ТЯ выносятся на обсуждение тюркологов и специалистов по инженерной лингвистике.

Литература

1. Севортян Э.В. Этимологический словарь тюркских языков. -М., 1974, 1978, 1980.

2. Мусаев К.М. Лексикология тюркских языков. - М.: Наука, 1984.

3. Кайдаров А.Т. Структура односложных корней и основ в казахском языке. - Алма-Ата: Наука, 1986.

4. Гузев В.Г., Пиотровский Р.Г., Щербак А.М. О создании машинного фонда тюркских языков // Сов. тюркология. - Баку, 1988. - № 2. - С. 92-101.

АЛГОРИТМИЗАЦИЯ ЗАДАЧИ ПОЛУЧЕНИЯ НА ЭВМ ФОНЕТИЧЕСКОЙ ТРАНСКРИПЦИИ КАЗАХСКОГО ТЕКСТА

АН Каз.ССР ИЯ им.Байтурсынова, Отдел фонетики.
Творческо-производственное объединение «ГРИФОН»
г.Ленинград. Тезисы докладов I Всесоюзная конференции
«Тюркская фонетика - 90», 26-28 ноября 1990 г.
Алма-Ата. – С. 87-88. (Соав.: С.А.Ускомбаев)

Известно, что в казахском языке, как и в других языках, буквы текста в силу определенных фонетических условий не всегда и не везде совпадают с обозначаемыми ими звуками. Следовательно, работы по определению частоты звуков, сочетаемости фонем и др., проведенные на основе текста, а таковые имеются, не отражают реальной ситуации в речевой деятельности человека. Для этого необходимо, чтобы исследуемый текст был фонетически транскрибирован.

Наш доклад освещает вопрос автоматического транскрибирования казахского текста в фонетическую запись с помощью ЭВМ. В этой области имеются опыты автоматического получения транскрипционной фонетической записи русских текстов. Однако алгоритмы, реализованные на ЭВМ для русского текста, оказались непригодными для казахского текста. Здесь вступают в силу такие принципиальные звуковые различия между русским и казахским языками, как флективность русского языка и агглютинативность казахского языка; словесное ударение в русском языке и сингармонизм в казахском языке, в русском языке имеется ряд согласных, которые реализуются только как мягкие или только как твердые, а в казахском языке все согласные имеют сингармонически твердые и мягкие, а также огубленные и неогубленные варианты. Кроме того, имеется ряд орфографических правил в обоих языках, которые требуют отличить решения друг от друга. С учетом указанных и др. особенностей казахского языка нами разработан алгоритм программы преобразования казахских текстов в фонетическую запись и написана программа на языке «Бейсик», который реализуется на ЭВМ «ЕС-1841». Результаты компьютерной обработки казахских текстов (протранскрибированный казахский текст) будет служить материалом для исследования различных звуковых закономерностей фонетики казахского языка: слогоделения, сочетаемости фонем, сегментации ревевого потока на уровне звука, слога и т.д. Думается, что разработанный нами алгоритм явится универсальным для всех тюркских языков.

О ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ С ПРИМЕНЕНИЕМ ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКИХ МЕТОДОВ В ТЮРКСКИХ ЯЗЫКАХ

АН Каз.ССР Институт языкознания. Труды группы «Автоматизация лингвистических исследовании и статистики речи», сб.: «Статистика казахского текста», выпуск II Алма-Ата, «Ғылым», 1990. 21-31-бб.
(Соавторы: А.Б.Белботаев, З.А.Сиразитдинов)

Последние двадцать лет в советском языкознании активно используются математические методы. Применение их не только в лингвистике, но и в социологии, психологии, археологии ң других гуманитарных науках свидетельствует о том, что эти науки вступают в новый качественный этап своего развития [1-3].

Среди разнообразных математических методов наиболее разработанным и часто применяемым методом в лингвистике является вероятностно-статистический метод. Этот метод позволяет выяснить те закономерности, которые действуют в языке и обусловлены многими массовыми процессами языка. Выяснить полностью их состав не представляется возможным.

Надо сказать, что отечественные и зарубежные ученые достаточно глубоко исследовали вопрос об использовании вероятностно-статистического метода.

Он успешно применяется при исследовании тюркских языков, таких, как казахский, узбекский, азербайджанский, киргизский, каракалпакский. Впервые этот метод был взят на вооружение казахстанскими языковедами. Для этой цели при Институте языкознания Академии наук Казахской ССР была создана группа статистико-лингвистических исследований и автоматизации. Она является неотъемлемой частью общесоюзной группы «Статистика речи».

Сейчас в Казахстане две группы математической лингвистики: чимкентская и алма-атинская. Они занимаются автоматизацией .лексикографических работ и статистическими исследованиями казахских текстов, а также проблемами машинного перевода. Итоги работы казахстанских статолингвистов обобщены в сборнике «Статистика казахского текста», изданном в 1973 г. [4]. В нем нашли отражение важнейшие выводы и рекомендации казахстанской школы структурно-прикладной и математической лингвистики. Так, приведены результаты опытов по составлению алфавитно-частотного словаря газет, сказок, языка художественной литературы, графемно-фонемной частотности казахских текстов, исследования по энтропии казахского языка и измерения смысловой информации слова. Впервые затрагивается проблема автоматизации и

моделирования не только казахских текстов, но и тюркских. Авторы предлагают машинные индексы для знаков орхоно-енисейской письменности, фонем каракалпакского и ногайского языков, звуков узбекского языка.

Многие важные вопросы, касающиеся вероятностно-статистического изучения языка и автоматизации этого процесса с применением средств вычислительной техники разработал К.Б.Бектаев. Его монография «Математические методы в языкознании» является фундаментальным пособием для всех лингвистов, использующих этот метод в своих работах [5].

Проблемы статистико-информационной типологии были впервые рассмотрены К.Б.Бектаевым в его монографии «Статистико-ішформационная типология тюркского текста» [6]. Она посвящена последовательному применению лингвистических приемов, вероятностных и информационных моделей к исследованию тюркского текста. Исследуя текст с помощью формальных статистико-вероятностных методов, автор описывает типологические особенности тюркских языков, сравнивая их с индоевропейскими языками. В Институте языкознания АН КазССР уделяется внимание новому направлению в исследовании тюркских языков. Так, в Алма-Ате был проведен Всесоюзный семинар «Статистическое и информационное изучение тюркских языков» (1969 г.)

Можно выделить основные исследования по фонетике и фонологии языка, слога и слоговой структуры слова, лексико-морфологической структуры текста и словоформ, а также функциональных стилей, синтаксиса текста, информационного характера языка.

Широко использовался статистический метод в работах по изучению лекепко-морфологической структуры современных литературных текстов. Главным компонентом такого изучения являются частотные словари: алфавитно-частотные и обратно-алфавитно-частотные словари основ слов и словоформ. Методика составления разработана достаточно. Проблема обсуждалась на многих всесоюзных конференциях, опубликованы подробные руководства по составлению частотных словарей и процессу автоматизации. Так, общие принципы индексации и кодирования современных тюркских текстов при автоматическом составлении с помощью ЭВМ были разработаны в Институте языкознания АН КазССР К.Б.Бектаевым и А.Х.Джубановым.

Первым частотным словарем в тюркском языкознании следует признать «Словарь языка Абая» [7]. Он представляет собой толковый словарь с указанием частоты употребления слов, который был составлен традиционным методом (вручную).

Сегодня имеются частотные словари, наиболее полно охватывающие язык, по казахскому и узбекскому языкам: частотные словари отдельных авторских произведений, языка публицистики, детской литературы, науки.

Если вести речь о словарях, то широко известный письменный памятник средневековья «Кодекс Кумапикус» в свое время был переведен на различные иностранные языки (латинский, немецкий, французский), однако

ни на один из языков народов СССР не переводился. Поэтому издание этого памятника на казахском языке с указанием частотных показателей стало заметным событием отечественной тюркологии [8].

Частотный словарь но роману М.О.Ауэзова «Абай жолы» («Путь Абая»), изданный в Алма-Ате в 1979 г. [9], явился началом для составления частотного словаря-словоуказателя языка произведений одного из основоположников казахской советской литератур; М.О.Ауэзова по полному собранию его сочинений в 20-ти томах. В этом лексикографическом труде найдут отражение лексическое и фразеологическое богатство языка писателя, индивидуальные особенности стилистических приемов словоупотребления и использования художественных и изобразительных средств казахского языка. Он дает богатый материал как для изучения особенностей языка писателя, так и для исследования статистической структуры казахской лексики вообще.

Объем исследуемых текстов произведений М.О.Ауэзова составляет более 2,5 млн словоупотреблений. В дальнейшем планируется также получить на ЭВМ распределительный (по томам) алфавитно-частотный словарь словоформ с указанием соответствующего номера тома и страниц, а также частоты употребления слов как по отдельным томам, так и в целом. Все тексты и словари-словоуказатели будут храниться на магнитных лентах.

Составлены и находятся на стадии завершения частотные словари каракалпакского и киргизского языков, исследуется язык публицистики. В 1988 г. увидел свет частотный словарь киргизского эпоса «Манас» К.Дыканова в двух частях.

Составление частотных словарей отдельных произведений авторов, функциональных стилей, всего литературного языка очень актуально. На их основе можно разрешить такой сложный вопрос, как. лексико-грамматическая норма языка. Эти словари могут быть использованы и при сравнительно-типологическом изучении тюркских языков, а также при создании машинного фонда тюркских языков. Начало такому исследованию уже положено.

Следует отметить, что составление частотных словарей важно не только в плане лексико-грамматического изучения языков, но и в методическом. Так, выяснение вида статистического распределения классов слов языка имеет большое значение при изучении «чисто» лингвистических явлений статистическим методом. Результаты анализа частотных словарей могут быть полезны при выяснении фундаментальных свойств и закономерностей конкретного языка, например параметров Ципфа для конкретного тюркского языка.

В плане нормативного изучения языка представляет интерес диссертационная работа А.Ахабаева, посвященная изучению лексико-морфологпческой структуры современной казахской публицистики [10]. В ней рассматривается, в частности, проблема статистического критерия нормативности одного из вариантов дублетных слов. На основе статистического анализа показано, что частотные характеристики могут

выступать в качестве такого объективного критерия. Посредством статисткко-морфологичсского анализа изучен вопрос о стилевой дифференциации вариантов грамматических форм казахского языка.

В целом же работа посвящена статистико-морфологическому анализу форм окончаний имен существительных и местоимений на основе сравнения газетных текстов за 1965-1966 гг. (150 тыс. словоупотреблений) и художественного текста (романа М.О.Ауэзова «Абай жолы», второй книги, 124 384 словоупотребления). Многие казахстанские ученые работают в этом направлении. А.Х.Джубанов сделал статистический анализ лексико-морфологических форм прилагательных по тексту романа М.Ауэзова «Абай жолы». Результаты анализа позволили выявить своеобразие языка автора в использовании класса слов – прилагательных казахского языка и их морфологических средств [11].

Если предыдущие авторы рассматривали статистико-морфологическую структуру имен казахского языка, то С.Мырзабековым исследована структура глагола современного казахского языка: основ и аффиксов, образующих производные формы глагола [12].

В диссертационной работе К.Молдабекова рассматривается проблема квантитативных особенностей лексики и морфологии языка казахской детской литературы и учебников начальных классов [13].

Статистические исследования казахских текстов проводятся и для выявления стилевых характеристик языка. Интерес представляют исследования А.Р.Зекеновой, которая на материале частотных словарей изучает язык пьес М.О.Ауэзова и других драматургов казахской литературы [14, 15].

А.Б.Белботаев на материале частотных словарей статистически изучил структурные формы окончаний казахского языка [16], а также проблему лингвистических терминов [17], исследовал лексику иноязычных слов, вошедших в произведения М.О.Ауэзова через русский язык [18, 19].

Проведены исследования количественной характеристики местоимений каракалпакского языка [20, 21]. А.Бабанаров составил частотный словарь языка современных турецких газет, исследовал лексико-статистическую структуру газетных текстов [22]. Ж.Жетешиков рассмотрел статистику словоизменительных аффиксов имен существительных киргизской публицистики, провел сравнительный анализ с данными казахского языка [23]. Статистический анализ использован С.Мухамедовым при изучении узбекских газетных текстов в его диссертационной работе [24].

М.Хусаинов статистический анализ использовал для оценки результатов, достигнутых реформой в области турецкой лексики [25]. С.Мамедовым изучены количественные характеристики словообразовательных элементов азербайджанского языка [26].

Т.М.Гарипов использовал статистический анализ для сравнительного лексико-семантического изучения тюркских языков. Результаты позволили выяснить, насколько такие языки, как башкирский, казахский, татарский,

чувашский, отличаются друг от друга как в плане количественном, так и в качественном [27].

Статистический метод с успехом применяется и при сравнительном изучении грамматических форм разных языков. Такие исследования проведены по казахскому и русскому, казахскому и английскому языкам [28, 29]. Так, А. Ергалауовым этот метод использован при изучении форм страдательного залога в английском и казахском языках [30].

Вероятностно-статистический метод в области фонетики применяется как при инструментально-экспериментальном подходе, так и при традиционном слуховом, при изучении как фонологической системы, так и графической [31, 32, 33].

Частотные характеристики графем, фонем, фонемосочетаний получены по казахскому, узбекскому, каракалпакскому языкам. В этой связи интерес представляют частотные словари сочетаний букв казахского языка, фонем каракалпакского языка, а также частотный словарь знакосочетаний Орхоно-Енисейского памятника [34].

В статье А.Х.Джубанова «К вопросу о графемной статистике казахского текста» сделана попытка статистического изучения казахского письменного текста на графемном уровне. Материал исследований может служить основой и примером для дальнейших работ, посвященных анализу морфемной структуры языка [35].

С.А.Ризаевым рассмотрен вопрос о частотности двухфонемных сочетаний в современном узбекском литературном языке [36]. По узбекскому и казахскому языкам изучены факторы, влияющие на относительную частотность фонем. Выяснено, что употребление фонем и их количественных отношений в разных жанрах различно. Сопоставлялись частотности фонем публицистического, научного и художественного по узбекскому и тексты сказок с научным по казахскому языкам. Статистический анализ позволил выделить группы фонем, определяющих специфику жанров.

С.С.Кенесбаевой применен статистический метод при определении наличия или отсутствия протетического звука в словах арабского происхождения, начинающихся со звуков, не характерных для исконно казахских слов [37]. А.И.Рабиновичем использована статистика для подтверждения теории о фонетической интерференции при контактах казахского и немецкого языков в условиях аудиторного билингвизма [38].

Проведены статистические анализы фонем турецкого и куманди-нского языков [39, 40].

А.Джунисбековым и Д.Маматовым проведены статистические анализы спектров фонем [41, 42]. Первым – гласных казахского языка, вторым – сложных согласных узбекского языка. По казахскому и узбекскому языкам проведены статистические исследования фонетической структуры языка: выявлены частотности индивидуальных фонем, позиционных вариантов. Начато такое исследование и по каракалпакскому языку [43].

Изучены комбинаторные свойства фонем азербайджанского языка [44].

По казахскому языку Д.А.Байтанаевой проведены статистико-информационные измерения всех фонем. Выяснено, что информация различна в разных жанрах языка [45].

С.С.Татубаевым исследованы статистические характеристики благозвучия певческой речи, в сравнении с разговорной речью. Выявлены количественные характеристики; определяющие благозвучие казахской певческой речи [46].

Вероятностно-статистический подход к изучению структуры слога и слоговой структуры слова применен в таких языках, как казахский, татарский, узбекский. К.Б.Бектаевым составлен частотный словарь слогов казахского языка на материале стилей художественных, научно-технических, публицистических текстов и двухтомного «Толкового словаря казахского языка», изданных в Алма-Ате в 1959, 1961 гг. В словаре указаны частота слога, по отношению его позиции в слове. Отдельно приводится словарь односложных слов. Д.А.Байтанаевой изучены статистико-информационные свойства слогов казахского письменного текста на основе статистической обработки опытно-экспериментальных данных [47, 48].

Т.И.Ибрагимовым выявлены состав и структурные типы слогов современного татарского литературного языка, определены комбинаторные свойства типов слогов, частоты употребления их в слове. Исследованы закономерности слоговой организации слова, в частности, выяснено, что предыдущий слог татарского языка несет 47% информации о последующем слоге, в значительной мере определяя его структурный и фонетический состав [49].

С.А.Ризаев исследовал статистическую структуру современного узбекского языка. Им выявлены типы слогов, изучены позиционные характеристики слога в слове, выяснены соотношения открытых и закрытых слогов, установлена каноническая форма слова современного узбекского подъязыка детской литературы [50].

Статистический анализ используется и при исследовании некоторых вопросов синтаксиса языка.

С.И.Ибрагимов на основе такого анализа текстов научно-технического характера выявил синтаксическо-функциональные классы существительных киргизского языка, определил виды синтаксических связей существительных с другими частями речи. Статистический анализ применен и при исследовании порядка слов в предложениях в казахском языке. Это позволило выявить особенности структуры простого предложения таких функциональных стилей, как литературно-художественный и научный. М.Р.Джумабаевой рассмотрены особенности порядка слов в простых предложениях английского и казахского языков [51, 52, 53].

В 1987 г. издана монография А.Х.Джубанова «Квантитативная структура казахского текста (опыт лингвистического анализа на ЭВМ)». В ней были рассмотрены опыт автоматического построения частотных списков лингвистических единиц тюркоязычных текстов, а также алгоритмы получения словаря-словоуказателя. Особо выделены автором вероятностно-

статистическая модель распределепня классов слов казахского языка и другие вопросы лингвистики [54].

В монографии Т.Садыкова «Проблемы моделирования тюркской морфологии (аспект порождения киргизской именной словоформы)» на основе системного подхода анализируется морфологическая структура киргизского языка, а результаты анализа используются в целях построения модели порождения реальных именных словоформ по заданной лексико-грамматической информации [55].

Приведенный нами обзор опубликованных работ но тюркским языкам, в которых применяется вероятностно-статистический метод, показывает, что круг исследований, в которых метод используется, довольно широк, охватывает все разделы языка. Это объясняется тем, что метод является универсальным, не зависящим от объекта исследования.

Однако не все анализируемые работы являются равноценными. Если по казахскому, киргизскому, азербайджанскому и узбекскому языкам проведены фундаментальные исследования па уровне диссертационных работ, то по другим языкам эти исследования только начаты, и о ходе работ можно судить только по кратким сообщениям и тезисам. Также неравномерно исследованы этим методом и сами области языка. Так, фундаментальные работы по выяснению статистической структуры в плане синтаксиса по какому-либо тюркскому языку еще отсутствуют. Это является отражением того, что статистико-вероятностный метод начал использоваться в исследованиях тюркских языков сравнительно недавно. Тем не менее он успел определиться как весьма полезный метод, и постепенно привлекает внимание ученых. Объясняется это прежде всего тем, что автоматизация трудоемких процессов в тюркологических исследованиях все еще желает лучшего. Проблема создания машинного фонда тюркских языков является жизненно важным делом и требует серьезного отношения со стороны тюркологов.

Литература

1. Количественные методы в социологии. М., 1966.

2. Количественные методы в гуманитарных науках. Минск, 1979.

3. Математические методы в историко-экономических и историко-культурных исследованиях. М., 1977.

4. Статистика казахского текста. Алма-Ата, 1973.

5. *Бектаев К. Б., Пиотровский Р. Г.* Математические методы в языкознании. Ч. 1: Теория вероятности и моделирование нормы языка. Алма-Ата, 1973. С. 282; *Бектаев К. Б., Пиотровский Р. Г.* Математические методы в языкознании Ч. 2: Математическая статистика и моделирование текста. Алма-Ата, 1974. С. 336.

6. *Бектаев К. Б.* Статистико-информационная типология тюркского текста. Алма-Ата, 1978. С. 183.

7. Абай тілі сөздігі. Алматы. 1968.

8. *Курышжанов А. К., Жубанов А. Х., Белботаев А. Б.* Куманско-казахский частотный словарь. Алма-Ата. 1978. С. 276.

9. *Бектаев К Б., Жұбанов А.Қ. Мырзабеков С., Белботаев* А. Б. М. Әуезовтің «Абай жолы» романының жиілік сөздігі (Частотный словарь по роману М.Ауэзова «Путь Абая»). Алматы, 1979. Б.334.

10. *Ахабаев А.* Статистический анализ лексико-морфолоогической структуры языка казахской публицистики. Алма-Ата, 1971. С. 24.

11. *Жубанов А. Х.* Статистическое исследование казахского текста с применением ЭВМ (на материале романа М. О. Ауэзова «Путь Абая»). Алма-Ата, 1973. С. 37.

12. *Мырзабеков С.* Статистико-лингвистический анализ структуры глагола современного казахского языка (по роману М. О. Ауэзова «Путь Абая»). Алма-Ата, 1973. С. 32.

13. *Молдабеков К.* Лингвистические исследования казахских текстов для младших школьников (на материале текстов учебников начальных классов)- Алма-Ата, 1985.

14. *Зекенова А.* М.О.Әуезов пьесаларындағы есім сөздердің морфологиялық құрылысы жайында. // Известия АН ҚазССР. Сер. филолог. 1976. С. 56–62.

15. *Зекенова А.* О лингвистическом исследовании драматургии М.О.Ауэзова//Материалы семинара «Статистическая оптимизация преподавания языков и инженерная лингвистика. Чимкент, 1980. С. 219–220.

16. *Белботаев А.Б.* Казақ тіліндегі жалғаулар және олардың құрылымдық формалары туралы // Материалы республиканской научно-теоретической конференции молодых ученых по общественным наукам. 24–25 февраля 1978 г. Алма-Ата, 1978. С. 181–183.

17. *Белботаев А. Б., Алдабергенова К. Қ.* Қазак тіл біліміндегі статолингвистикалық терминдер мен сөз тіркестері туралы//Материалы республиканской научной конференции молодых ученых и специалистов по общественным наукам, посвященной 110-й годовщине, со для рождения В. И. Ленина (17–18 апреля 1980 г.). Алма-Ата, 1981. С. 178–179.

18. *Алдабергенова Қ. Қ., Белботаев А.* Статистическое исследование морфологической структуры иноязычных слов, вошедших в произведения М. Ауэзова через русский язык // Материалы семинара «Статистическая оптимизация преподавания языков и инженерная лингвистика». Чимкент, 1980. С. 149–150.

19. *Алдабергенова Қ. Қ., Белботаев А. Б.* Количественно-морфологический анализ слов из русского языка, вошедших в рассказы М. О. Ауэзова // Вопросы тюркского языкознания. Алма-Ата, 1985. С. 164–166.

20. *Бектаев К. Б., Насыров Д. С., Аимбстов М. К.* Количественная характеристика личных и лично-притяжательных местоимений каракалпакского языка // Вестник Каракалпакского филиала АН УзССР. 1982. № 1. С. 67—71.

21. *Насыров Д. С, Бектаев* /(. *Б., Аимбетов М. I*\. Количественная характеристика указательных и указательно-притяжательных местоимений

каракалпакского языка, употребляемых в публицистике // Вестник Каракалпакского филиала АН УзССР. 1981. № 4. С. 79–82.

22. *Бабанаров А.* Частотный словник и автоматический словарь для машинного перевода турецких газетных текстов // Инженерная лингвистика и оптимизация преподавания иностранных языков. Л., 1980.

23. *Жетеишков Ж.* Статистический анализ словоизменительных аффиксов имен существительных киргизского языка (на материале газетных текстов) // Материалы семинара «Статистическая оптимизация преподавания языков и инженерная лингвистика». Чимкент, 1980. С. 210–211.

24. *Мухамедов С.* Статистический анализ лекеико-морфологической структуры узбекских газетных текстов. Ташкент, 1980. С. 25.

25. *Хусаинов С.* Из опыта статистического анализа частотности употребления неологизмов в современном турецком литературном языке // Советская тюркология. 1981. № 4. С. 74–79.

26. *Мамедов Э. С.* Количественные характеристики словообразовательных элементов // Научные труды Азербайджанского пединститута рус. яз. и лит-ры. Сер. яз. и лит. Баку, 1979. № 1. С. 105–110.

27. *Гарипов Т. М.* Лексикостатистические отношения тюркских языков. Урало-Поволжья // Языковые контакты в Башкирии. Уфа, 1972, С. 248–267.

28. *Ишанов К. И., Садиков А. В., Мырзабеков С.* Частотно-сопоставительная характеристика падежных форм русского и казахского языков // Тезисы докладов и сообщений всесоюзного семинара «Статистическое и информационное изучение тюркских языков». Алма-Ата, 1969. С. 56–58,

29. *Ахметов Е. Г.* Статистическая характеристика родительного падежа имени существительного в английском и казахском языках // Тезисы докладов и сообщений всесоюзного семинара «Статистическое и информационное изучение тюркских языков». Алма-Ата, 1969. С. 51–54.

30. *Ергалауов А.* Статистическая характеристика форм страдательного залога в английском и казахском языках // Тезисы докладов и сообщений всесоюзного семинара «Статистическое и информационное изучение тюркских языков», Алма-Ата, 1969. С. 62–63.

31. *Джубанов А. Х.* К вопросу о графемной статистике казахского текста // Вопросы казахской фонетики и фонологии. Алма-Ата, 1979. С. 79–86.

32. *Джубанов А. Х., Джунисбеков А.* Преобразование с помощью ЭВМ казахских текстов в фонетическую запись // Материалы семинара «Статистическая оптимизация преподавания языков и инженерная лингвистика». Чимкент, 1980. С. 317–319.

33. *Исенгельдина А. А.* Факторы, определяющие относительную частотность фонем // Статистика казахского языка. Алма-Ата, 1973. С. 659–664.

34. Статистика казахского текста. Алма-Ата, 1973.

35. *Джубанов А. Х.* К. вопросу о графемной статистике казахского текста // Вопросы казахской фонетики и фонологии. Алма-Ата, 1979. С. 79–

86.

36. *Ризаев С. А.* К вопросу о частотности двухфонемных сочетаний в современном узбекском литературном языке // Опыт экспериментального и структурного изучения языка. Ташкент, 1982. С. 45–92.

37. *Кенесбаева С. С.* К вопросу исследования протетического звука в словах арабского происхождения статистическим методом // Тезисы докладов и сообщений всесоюзного семинара «Статистические и информационные изучения тюркских языков». Алма-Ата, 1969. С. 43–44.

38. *Рабинович А. И.* О статистическом исследовании казахско-немецкой фонетической интерференции // Там же. С. 65—66.

39. *Чикадзе В. Г.* Статистический анализ фонем в односложных словах турецкого языка // Сообщения АН ГССР. Тбилиси. Т. 96. № 1. С. 201–209.

40. *Селютина И. Я.* Относительная частотность согласных в языке кумандинцев // Экспериментально-фонетические исследования сибирских языков. Новосибирск, 1981. С. 76—92.

41. *Джунисбеков А.* Гласные казахского языка. Алма-Ата, 1972. С. 94.

42. *Маматов Д.* Количественный анализ сложных согласных//Опыт экспериментального и структурного изучения языка. Ташкент, 1982. С. 108–111.

43. *Насыров Д. Х., Аимбетов А.* О статистике фонем каракалпакского языка // Материалы семинара «Статистическая оптимизация преподавания языков и инженерная лингвистика». Чимкент, 1980. С. 273–274.

44. *Вейсалов Ф. Е., Керимов Я. К., Сафаров М. Г.* Комбинаторика фонем современного азербайджанского языка // Вопросы азербайджанского языкознания. Баку, 1981.

45. *Байтанаева Д. А.* Информационные характеристики казахского текста. Алма-Ата, 1985. С. 18.

46. *Татубаев С. С.* Статистические подходы к исследованию певческой фонетики казахского языка // Тезисы докладов и сообщений всесоюзного семинара «Статистическое и информационное изучение тюркских языков». Алма-Ата. 1969, С. 38–40.

47. *Бектаев К. Б.* Алфавитно-частотный словарь слогов казахского языка // Статистика казахского текста. Алма-Ата, 1973. С. 566–612.

48. *Байтанаева Д. А.* Энтропия распределения слогов казахского письменного текста // Тезисы докладов и сообщений всесоюзного семинара «Статистическое и информационное изучение тюркских языков». Алма-Ата, 1969. С. 80–82.

49. *Ибрагимов Т. И.* Некоторые статистические данные о слогах татарского языка: вероятностные методы и кибернетика // Учен. зап. Казанского ун-та. Вып. 4. Т. 25. Кн. 6. 1966. С. 74–78; *Его же.* Изучение слогов и структуры их сочетаний в татарском литературном языке. Казань, 1970.

50. *Ризаев С. А.* Из опыта исследования статистической структуры слога в узбекском языке // Общественные науки в Узбекистане. Ташкент, 1972. №1; *Его же.* Статолингвистический анализ структуры слова в

современном узбекском языке // Материалы семинара «Статистическая оптимизация преподавания языков и инженерная лингвистика». Алма-Ата, 1980. С. 180–182.

51. *Ибрагимов С. И.* Некоторые статистические характеристики имен существительных киргизского языка // Тезисы докладов и сообщений всесоюзного семинара «Статистическое и информационное изучение тюркских языков». Алма-Ата, 1969. С. 34–38.

52. *Ахабаев А., Белботаев А., Молдабеков К., Осижигитова Н.* Статистика порядка слов в предложении // Там же.

53. *Джумабаева М. Р.* О статистической характеристике, простого предложения в английском и казахском языках // Там же. С. 62—63.

54. *Джубанов А. Х.* Квантитативная структура казахского текста (опыт лингвистического анализа на ЭВМ). Алма-Ата, 1987. С. 147.

55. *Садыков Т.* Проблемы моделирования тюркской морфологии (аспект порождения киргизской именной словоформы). Фрунзе. 1987. С, 121.

ИСПОЛЬЗОВАНИЕ ЭВМ ДЛЯ ВЫДЕЛЕНИЯ ОПОРНЫХ СЛОВ ТЕКСТА

Тартуский Университет «Актуальные проблемы компьютерной лингвистики». Тезисы докладов Всесоюзной конференции в г.Тарту 29-31 мая 1990 г. – С. 49-50.

Проблемы выделения основного и факультативного содержания текста является одной из важных проблем инженерной лингвистики. Существует много различных способов выделения опорных (ключевых) слов. Наиболее перспективным нам представляется метод использованный в работах Марусенко М.А. и Зубова А.В. [1; 2]. В его основу положены идеи о том, что наиболее значимые единицы текста имеют наибольшую частоту употребления в тексте и встречаются в наибольшем количестве абзацев такого текста. Формально эта зависимость выражается в виде следующей формулы:

$$\text{Кваж} = \frac{F \cdot n}{N \cdot n},$$

где F – абсолютная частота слова в тексте, m – число абзацев, в которых встретилось слово, N – общее число слов в тексте, n – общее число абзацев в тексте.

Как видно, подсчеты по этой формуле требуют построения частотного словаря как отдельного текста, так и всех его абзацев. Ясно, что для достаточно больших текстов эта задача трудно выполнима.

Поэтому нами создана программа, которая строит частотные словари абзацев и в совокупности частотный словарь всего текста. Предварительно перед вводом текста в ЭВМ все словоформы текста приводятся к канонической (словарной) форме. Затем машина строит частотные словари

лексем каждого абзаца. И в заключении, компьютер, объединяя подобные словари абзацев строит частотно-алфавитный словарь всего текста. При этом для каждой лексемы словаря текста автоматически находится абсолютная частота лексемы в тексте (F, см. формулу (I)) и число абзацев, в которых встретилась данная лексема (m). Все эти данные используются затем для автоматического отнесения слов анализируемого текста к числу главных или второстепенных опорных слов.

В докладе подробно на материале казахских текстов рассматривается процедура построения словарей абзацев и всего текста.

Литература

1. Зубов А.В. Статистический аспект содержания текста и его формальное представление // Квантитативная лингвистика и автоматический анализ текстов. – Тарту, 1986, с. 75-94.

2. Марусенко М.А. О формировании словника словаря статистически устойчивых научно-технических терминов // Структурная и прикладная лингвистика. – Л.: ЛГУ, 1983, с. 82-89.

АВТОМАТИЧЕСКИЙ ОТБОР ОПОРНЫХ СЛОВ ТЕКСТА В ЦЕЛЯХ ОПТИМИЗАЦИИ ОБУЧЕНИЯ КАЗАХСКОМУ ЯЗЫКУ

Госкомитет по нар-у образ-нию СССР.
Мин-во нар-го образ-ния БССР. Минский
Гос. пед. институт иностраннных языков.
Тезисы докл-в Всесоюзной конференции:
«Использование ЭВМ в научной и учебной
работе гуманитарного ВУЗА». Минск,
12-14 июня 1990 г. – С. 67-68.

В связи о принятием закона о языках а Казахской ССР большое значение придается рациональным методам обучения казахского, русского и других языков народов, проживающих в республике.

Одним из условий успешного усвоения любого национального языка, в частности казахского, является умелое и научно-обоснованное применение методов информатики и использование в процессе обучения возможностей современных компьютеров.

Эффективность обучения казахскому языку во многом определяется тем минимумом лексики, которым владеет обучаемый. А этот минимум зависит от большого количества факторов: образование обучаемого, той литературы, которую читает обучаемый, образовательного уровня родителей и т.п. и т.д.

Наиболее ярко эти интересы обучаемых преломляются в так называемых функциональных языковых стилях: художественной литературе, драме, поэзии, научной литературе, публицистике.

Поэтому возникает задача построения частотных словарей, принадлежащих соответствующим стилям. Надежные данные при этом могут быть получены в том случае, если анализу будут подвергаться тексты значительных объемов. Очевидно, что без применения современных компьютеров такая задача трудно выполнима.

В процессе построения формальной модели казахского языка нами создан комплекс программ, который позволяет строить на машинах серий ЕС–ЭВМ частотные словари по текстам указанных пяти стилей.

Однако психолингвистическими экспериментами установлено, что то или иное казахское слово лучше запоминается в процессе обучения в зависимости от того, насколько часто оно употреблено и в каком количестве абзацев и текстов. Это приводит к необходимости построения не только простых частотных словарей текстов, но и построения так называемых частотно-распределительных словарей, где при каждом слове указывается не только абсолютная частота употребления слова, но и число абзацев и текстов, в которых данное слово употреблено в пределах рассматриваемого стиля.

Те слова текстов, которые имеют наибольшую абсолютную частоту и встречаются в наибольшем числа абзацев и текстов, должны быть даны для обучения этому стилю в первую очередь.

В докладе детально рассматривается указанная процедура построения частотно-распределительного словаря по отношению к стилям казахской художественной литературы и публицистики.

АВТОМАТИЧЕСКИЙ АНАЛИЗ МОРФОЛОГИИ КАЗАХСКОЙ СЛОВОФОРМЫ

АН Каз.ССР Институт языкознания. Труды группы
«Автоматизация лингвистических исследовании
и статистики речи», сб.: «Статистика казахского текста»,
выпуск II Алма-Ата, «Ғылым», 1990. 148-156-бб.
(Соавтор – С. А. Ускомбаев)

Автоматический анализ морфологии слова является одним из наиболее важных и необходимых этапов лингвистической операции. Путем автоматического морфологического разбора при обработке отдельно взятых словоформ обеспечивается получение первичных грамматических характеристик и отделение морфемной основы словоформ от аффиксальных морфем. Это необходимо для последующих этапов распознавания словесной величины.

В системах автоматической обработки текстов (например, в автоматической поисковой системе) морфологический анализ применяется для установления фактов парадигматических соотносительностей слов и выявления связующих грамматических показателей слов и их сочетаний, необходимых для исследования на синтаксическом уровне. Опыт

моделирования лексики и морфологии некоторых индоевропейских языков [1–3] показал, что в связи с инженерно-лингвистическим моделированием и автоматическим переводом текстов возникает необходимость создания машинной морфологии, «опирающихся на нетрадиционное выделение так называемых машинных флексий и создание словаря машинных основ (квазиоснов)» [4, 5]. В результате автоматического анализа морфемного состава можно получить первичные, грамматические признаки только по предписанному грамматическому коду.

Автоматическое приведение словоформы к основной (словарной) форме является сложной задачей из-за большого количества фоно-морфологических особенностей лексики, представляющих национальные черты тюркских языков, в частности казахского.

Задачи автоматического морфологического анализа казахских текстов выдвигают необходимость накопления и хранения максимально полной информации об отдельном слове (графические, фонетические характеристики и даже семантические, парадигматические и социолингвистические).

Наши исследования, направленные для решения задачи лемматизацнп, проводятся па материалах десятитомного толкового словаря казахского языка, алфавптно-частотиых, обратных словарей, составленных на ЭВМ по отдельным жанрам казахских письменных текстов различного объема. Изучение и выявление внутреннего строения казахской словоформы и всевозможных правил порождения с помощью аффиксов и основ казахских слов, является одной из важных задач при автоматическом морфологическом анализе казахского текста. На базе этих исследований разрабатываются алгоритмы автоматического слогоделения и автоматического морфологического анализа. Ранее полученные А. Ахабаевым, С. Мырзабековым и А. Джубановым статистические данные по морфологической структуре казахских существительных, глаголов, прилагательных, местоимений [6] послужили материалом для выявления и изучения закономерностей аффиксации словоформ по отдельным частям речи. В алгоритме автоматического морфологического анализа мы попытались учесть случаи соответствия словоформ, соотносящихся с основной формой слова, при котором возникает немало трудностей. Например, слова «бал» (мёд), «балға» (молоток) в словосочетаниях «балға мал» (мокни в мёд) и «үлкен балға» (большой молоток) затрудняют автоматическое выделение основы «бал» или «балға». Поэтому такие омонимичные слова задаются как исключения в виде отдельной таблицы. Для ускорения скорости поиска в нашем алгоритме также указывается длина основы в слогах, намного упрощающая процесс отделения основ от казахских словоформ. Кроме словаря основ и словоформ, в алгоритме используется также словарь сочетаний словообразующих и формообразующих аффиксов. При этом аффиксы, присоединяемые к существительным и образующие существительные от других классов слов, составляют одну группу, а аффиксы, присоединяемые к существительным и

образующие глаголы, – в другую группу и т. д. Словарь сочетаний аффиксов можно использовать и для установления парадигматических эквивалентов слов.

В казахском языке только у существительных количество формообразующих аффиксов достигает 500, а многие казахские глаголы имеют до 1000 различных форм [7]. В связи с этим возникает необходимость проводить классификацию частей речи и грамматических категорий путем машинной обработки морфологических данных для их формализации.

В тюркских языках части речи традиционно классифицируются в основном с точки зрения их смыслового содержания слова (лексического), т.е. значения принадлежности к той пли иной системе словообразования или словоизменения, а также синтаксических функций.

На основе этих признаков в тюркских языках первичные и вторичные морфемы делятся на три группы [8]:

1. Корневые морфемы, выражающие имя, т. е. название предмета или статического признака предмета;

2. Корневые морфемы, выражающие глагол, т. е. обозначающие действие – динамический признак;

3. Корневые служебные морфемы, выражающие граммагические отношения слов в составе словосочетаний и предложений.

К первой группе относятся имена существительные, имена прилагательные, наречия, числительные, междометия и мимемы, местоимения, ко второй группе – глаголы и к третьей – частицы, послелоги, союзы, модальные слова.

Слова, относящиеся, к первым двум группам, имеют особые, отличные друг от друга системы аффиксального словообразования. Однако формально некоторые омоаффиксы неразличимы (омографы). Омонимичные слова могут иметь аффйксы, присоединяемые к корневым морфемам как первого, так и второго типа. Всю совокупность корневых морфем можно подразделить также на два больших класса: самостоятельные, или знаменательные; несамостоятельные, или служебные.

Кроме того, в тюркских языках существуют группы слов, которые на морфологическом уровне, т.е. вне контекста, нельзя отнести к определенной части речи (омонимичные слова).

Для изучения природы казахского языка необходимо хорошо представлять систему слоговых структур. Казахская письменность, как и многие письменности индоевропейских языков, основана на звуковой системе записи. Несмотря на это, в казахском языке больше всего преобладает слоговой оттенок. Поэтому знания морфологической структуры языка сопряжены с умением сегментировать слова казахского языка на слоги. Еще в 1934 г. Х. Жубанов в статье «Буын жігін қалай табуға болады?» (Как определить границы слогов? [9] на базе выделенных им шести свойств слогов предложил своеобразный алгоритм определения границ слогов в казахских словах (словоформах). Ниже перечислены шесть свойств, характерных для казахского языка, когда:

– не существует слога без участия гласного звука;

– в одном слоге не может быть более одного гласного;

– изолированный гласный звук может, образовать слог и самостоятельно;

– в одном слоге участвует не более трех согласных;

– сочетание двух согласных в одном слоге возможно только после гласного звука; причем в обязательной композите: сонор (звонкий)+глухой;

– внутри слова не встречается слог, начинающийся с гласного звука. В середине или конце слова слог обязательно начинается с инициального звука.

Сегментацию слов на слоги в казахском в языке профессор Х. Жубанов предлагает начать с конца слова, причем конечному гласному должен предшествовать согласный звук, который и означает начало границ между слогами.

Количество звуков в слогах собственно казахских словоформ определяется в пределах от одного до четырех.

Х. Жубанов выделяет шесть вариантов слогов по разновидности структур звуков: Г; СГ; ГС; СГС; СГСС.

Основываясь на этом, построим алгоритм слогоделения казахской словоформы, блок схемы его ниже (см: блок-схему 1).

Автоматический морфологический анализ казахской словоформы, а именно приведение любой словоформы текста к реестровой форме мы проводим согласно слоговой структуре казахского слова.

В блок-схеме 2 (здесь из-за объема книги данную блок-схему мы решили не приводить см. сб.: «Статистика казахского текста», выпуск II Алма-Ата, «Ғылым», 1990. 150-152 стр. и 154-155 стр.) в укрупненном виде приводится алгоритм вышеназванной задачи. Решение ее основано на распознавании словообразующих и словоизменяющих аффиксов среди слоговой структуры конкретного слога. Анализ проводится с конечных слогов слова, путем сравнения с заранее заданным перечнем названных аффиксов, производится отсечение конечных словоизменяющих аффиксов. При этом учитываются случаи омонимии аффиксов. Некоторые конкретные ситуации заданы в виде таблиц исключений. Те случаи, которые не подчиняются нашим правилам сегментации, выдаются на экран дисплея для последующей визуальной корректировки. Такие нетипичные случаи фиксируются в таблице исключений. Данный алгоритм реализован на разных, различающихся по стилю и типам, текстах.

Литература

1. Статистика речи. Л., 1969. С. 259.

2. Статистика речи и автоматический анализ. Л., 1971 г. С. 240.

3.*Зореф М. Г.* Машинная морфология и ее применение в немецко-русском автоматическом словаре // СТРААТ. Л., 1973. С. 270.

4. *Коваль Н. А.* Опыт построения словаря основ и таблиц окончания украинского глагола для анализа при машинном переводе // Применение новых методов в изучении языка. Вып. 1. Днепропетровск, 1969. С. 245.

5. *Баскаков Н. А.* Морфологическая структура слова и части речи в тюркских языках // Советское востоковедение. 1957. № 1. С. 72–85.

6. Статистика казахского текста. Алма-Ата, 1973. С. 49–136, 231.

7. Там же. С. 39.

8.Там же.

9. *Жубанов Х.* Исследования по казахскому языку. Алма-Ата, 1965.

ТҮРКІ ТІЛДЕРІНІҢ МАШИНАЛЫҚ ҚОРЫН ЖАСАУДЫҢ АЛҒЫШАРТТАРЫ

Қаз. ССР ҒА Тіл білімі институты. «Лингвистикалық жұмыстарды автоматтандыру және тіл статистикасы» тобы қызметкерлерінің еңбектері: «Қазақ тексінің статистикасы». II шығарылуы. Алматы, «Ғылым», 1990, 31-42-бб.

Электронды техниканың дүниеге келуі қазіргі кезде инженерлік-экономикалық міндеттерді шешумен бірге тіл білімі саласында да өз көмегін тигізуде. Мысалы, электронды есептеу машиналары арқылы бір тілден екінші тілге аударуға, ғылыми-техникалық текстердің рефераттары мен аннотацияларын автоматты жолмен жасауға, қажетті информацияларды жедел тауып алуға кең жол ашылуда. Электронды есептеу машиналар (бұдан әрі – ЭВМ) бұл күндері практикалық тіл білімі саласындағы әсіресе еңбекті көп тілейтін операцияларды ішінара автоматтандыруға, текст түріндегі материалдарды әралуан белгілеріне қарай сұрыптауға жиі пайдаланылып отыр.

Электрониканың өркендеуі, дәл методтардың қолданыс табуы тіл білімінің сан алуан салаларын, соның ішінде, лингвистика ғылымын да жеделдете дамытуға ықпалын тигізуде. Үнді-европа тіл білімі мамандары қолданған математикалық және кибернетикалық методтардың әсері түркітану бағытындағы зерттеулерге де тимей қойған жоқ. Алайда бұл құбылыс 60 жылдарға дейін айтарлықтай еленіп, ескерілген жоқ.

Біз түркі тілдерін ЭВМ-ді қолдана отырып санамалы (количественный) метод жолымен зерттеудің белгілі бір теориялық мәні барлығын атап айтуымыз қажет. Өйткені үнді-европаның қопармалы (флективті) аналитикалық тілдері бойынша жинақталған ұлан-ғайыр статистикалық материалдарды жалғамалы (агглютинативті) түркі тілдерінің соңғы статистикалық деректерімен салыстырудың мүмкіншілігі артады. Зерттеулеріміздің нәтижелеріне қарағанда, түркі тілдерінің табиғаты сөйленіс ағымында жекелеген сөз тұлғаларының түбегейлі фонетикалық және морфологиялық өзгерістерге ұшырауын қажет етпейтіндігі, олар негізінен өздерінің бітім-бейнелерін сол қалпында сақтайтындығы

айқындалып отыр. Дыбыстардың үнді-европа тілдеріндегі толассыз тоғысулары, әлсіреп редукциялануы тағы басқа қасиеттер түркі тілдеріне тән емес. Ана тіліміздің осы тәрізді ерекшеліктеріне сүйеніп, оның алуан сырлы құбылыстарын математикалық жолмен зерттеуге дәл және даусыз хаттауға түркі тіл білімінің бірсыпыра процестерін автоматтандырудың тигізер пайдасы мол.

Лингвистикалық зерттеулерді автоматтандыру кезінде «адам-машина-адам» жүйесі бойынша жұмыс істеледі. Сондықтан адам мен ЭВМ-нің қызметіндегі кейбір ұқсастықтар мен айырмашылықтар мыналар: машина (ЭВМ) шапшаң әрекет жасайды және барлық амалдарын дәлме-дәл орындайды, оның жадындағы информациялар ұзақ мерзім бойында сақталады, әрі қажетті сәтінде қолма-қол іздеп табуға мүмкіншілік береді. Ал адамның ми қызметінің ерекшелігі – эвристикалық немесе тапқыштық қабілетінде (мидың қызметінің басқа қабілетін айтпағанның өзінде). Сонымен лингвистикалық зерттеулерді автоматтандыру жұмысын ЭВМ мен адамның ойлау қабілетіндегі ерекше сипаттарды ескере отырып жүзеге асыруға тиіспіз. Шетелдердегі, сондай-ақ Совет Одағындағы ғылыми топтардың қазіргі жұмыс тәжірибелерінің негізінде мынадай тоқтамға келуге болады: ЭВМ-нің көмегімен тек қолданбалы тіл білімі ғана емес, тіпті басқа да тіл зерттеу саласындағы өзіндік ішкі теориялық проблемаларды шешуде де қомақты нәтижелерге қол жететіні даусыз. Яғни электронды машиналардың көмегімен лексиканы, грамматиканы, фонетиканы, қазіргі және көне жазба текстердің басқа да сипаттарын хатқа түсіруде оның берері мол. Лингвистикалық процестерді автоматтандыру жұмысы ғалымдардың тұжырымы бойынша екі деңгейде болмақ: а) текстерді комбинаторлы-статистикалық әдіспен суреттеп шығу, ә) текстерді семантикалық шешімдер қабылдау тұрғысынан қарастыру.

Лингвистикалық зерттеулерде ЭВМ-ді қолдану көбіне әріптердің, фонемалардың, сөздердің, сөз тұлғаларының, сөз тіркестерінің т.б. жиілік тізбелерін (сөздіктерін) жасаудан басталады. Әр дәуірдегі түрлі авторлардың шығармаларының текстеріне сәйкес шығарылған жиілік көрсеткіштерінің арқасында қазіргі кезде жалпы лексикологияның мәселелері жалпыхалықтық лексиканың құрылымы оның әр алуан қабаттарының арақатынасы, қазіргі замандағы тілдік норма, жазушы тілінің функционалды стилі мен ерекшелігі, сөз қорының тарихи дамуы т. б. аса маңызды проблемалар принциптік тұрғыдан жаңаша пайымдалып отыр.

Тек ЭВМ-ді қолданудың арқасында ғана кері алфавитті жиілік сөздігі тәрізді құралдарды шығару мүмкін болды. Мұндай сөздік әрбір сөздің грамматикалық сипатын түсіндіруге аса оңтайлы болғандықтан морфология саласындағы түрлі зерттеу жұмыстарының міндеттерін анағұрлым жеңілдетеді. Мәселен, оны белгілі бір тілдегі түрлендіруші және сөз жасаушы тәсілдердің толық сипатын айқындауға санамалы қатынастарды білуге, сондай-ақ фонетика мен фонология өрісіндегі толып жатқан мәселелерді шешуге пайдалануға мейлінше болады. Өйткені, кері алфавитті сөздікте жазылуын реттестіретін сөздердің жиыны бесенеден белгілі болып

көрініп тұрады. Қазірде ЭВМ арқылы әралуан жиілік сөздіктерді жасаудың тәжірибелері бірсыпыра жинақталды. Мұндай сөздіктер түрліше мақсатпен және әркелкі көлемде үнді-европа тілдері бойынша да сондай-ақ түркі тілдері бойынша да көптеп жасалуда.

ЭВМ-ді лингвистикалық немесе жалпы филологиялық мақсаттардағы зерттеулерге пайдаланумен бірге сол ғылыми ізденістердің нәтижелерін іске асыру мақсатында да қолдануға болады. Мұның өзі ғылыми-техникалық процестерді жеделдету туралы КПСС XXVII съезі алға қойған міндеттерге сай ғылыми-зерттеу істерін және қазіргі өндірістік процестерді кеңінен автоматтандыру туралы талаптарға барынша үндесіп жатады. Осылайша электронды машиналарды пайдалана отырып тілдерді зерттеу және ғылыми қорытындыларды табиғи тілдердің информациясын автоматты жолмен танып-білу мүмкіндігіне саналы түрде көз жеткізудің арқасында ғалымдар орыс тілінің машиналық қорын жасау қажеттігі жөніндегі идеяларды жүзеге асыра бастады. Осы мақсатпен 1983 жылы шақырылған бірінші және 1987 жылы шақырылған екінші Бүкілодақтық конференцияларда орыс тілінің машиналық қорын жасау проблемалары күн тәртібіне қойылып, онда машинаға түсірілген тілдің құрамы мен құрылымы қандай болуы тиіс екендігі, бірінші кезекте мұндай тіл қандай міндеттер атқаратыны, машина қоры қамтитын тіл көздерінің ауқымы қандай болмақтығы (минимум және максимум мөлшерде), машиналық тіл қорын дәйекті түрде тұтыну үшін қандай ұйымдық және техникалық шаралар атқарылуы керектігі талқыланды.

Осы программаны жүзеге асыруға, «қағазсыз информацияның» технологиялық тәсілдерін игеруге, сөйтіп өз еңбегінің өнімділігін арттыруға академиялық ғылыми мекемелер мен жоғары оқу орындарындағы тіл мамандарының мейлінше ынталы екендігін конференция айқын танытты.

Егер алғашқы конференцияда орыс тілінің машиналық қоры кандай болуы қажет деген ділгерлігі күшті проблемалар талқыланған болса, екінші конференцияда осы салада орындалған жұмыстардың нәтижелері сөз болды және жаңа саладағы жұмыс барысында орыс тілінің және басқа ұлттар тілдерінің машиналық қорларын жасау мен байланысты күн тәртібіне тосыннан қойылған соны проблемалар сөз болды. Атап айтқанда СССР халықтары тілдерінің машиналық қорларын жасау мәселесі өз алдына жеке тақырып ретінде қаралды. Ерекше атап кетерлік бір жайт мынау болды: конференцияда орыс, украин, грузин, эстон тілдері жөніндегі баяндамалармен қатар түркі тілдерінің, оның ішінде қазақ тілінің машиналық қорын жасау мәселесі жеке тыңдалды. Сөйтіп, қазақ текстерін ЭВМ-ге енгізу және автоматты жолмен зерттеу бағытындағы біздің көпжылдық тәжірибелеріміз бен қол жеткен нәтижелеріміз ескеріліп, екінші конференцияның шешімдерінде түркі тілдерінің машиналық қорын жасайтын координациялық орталық Қазақ ССР ҒА Тіл білімі институтының құрамынан ашылсын деген дербес қаулы қабылданды.

Біз осы мақаламызда тілдің машиналық қорының мән-жағдайы жайлы өз ойларын ортаға салған ірі ғалымдардың жоғарыда сөз болған ғылыми форумдағы кейбір тұжырымдарына тоқтала кетпекпіз.

Машиналық қор жайындағы мәселені ең алғаш рет белгілі совет ғалымы, информатика саласының маманы А.П. Ершов 1978 жылы «Іс қағаздары прозасының феномены: диалогтар жүйесін жасау методологиясына» атты мақаласында сөз етті[1]. Бұл мәселенің ойдағыдай шешілуіне ғалым үлкен мән беріп, егер орыс тілінің машиналық қорын жасау мақсаты ойдағыдай орындалса, модельдеу мен алгоритмдерді құрастыру саласындағы қол жеткен табыстарымыздың бәрі академиялық жаттығулар қалпында қалып қоймай, іс жүзіне асқан болар еді, мұндай өте-мөте қиын проблеманың шешілуі ғылым, мәдениет және қолданбалы техникалық тәсілдер тұрғысынан мейлінше құнды болмақ деген болатын. Оның пікірінше, маман лингвистер тарапынан орыс тілінің машиналық қорын жасау ісі өндірістік лингвистикалық жүйелерді іске қосудан озық болуы тиіс, өйткені тек осы жағдайда ғана толып жатқан қажетсіз қайталаулардан құтылып, тілдің жанды өзегін озбырлар мен дүмшелердің қиянатынан аман сақтап қалуға болады.

«Орыс тілін сүйетін қоғамның (формалды қоғам) танымынша, дейді ол, орыс тіл білімі дегеніміз әзірге ұйықтап жатқан Гулливер тәрізді, сондықтан оның оянатын және өзі туралы бар дауыспен жар салатын кезі енді жетті. Ал, ЭВМ бұл ретте таптырмайтын қоңыраулы сағат секілді», - дейді ғалым. А.П. Ершов болашақ жасалатын орыс тілінің машиналық қорын тіл жөніндегі білімдердің аса қуатты концентраты болады деп сенеді. Мұндай формалды жүйе табиғи тілдің организмімен сәйкес әрі тең көлемді болуы қажет, сөйте тұра ол (модель) анатомиялық тұрғыдан мүшеленген, ашық, айқын бақылауға, зерттеуге қолайлы әрі өзгермелі болуы тиіс. Алайда, қазіргі жансыз материяның формаларын модельдеуге қатысты білімдеріміздің аясын осы күнгі тіл білімінің статусымен және ауқымымен салыстырып қарайтын болсақ, біз аталған салада жүріп өтпек жолымыздың бастауында тұрғанымыз айқын сезіледі.

Ю.Н. Карауловтың «Лингвистикалық зерттеулердің методологиясы және орыс тілінің машиналық қоры» атты баяндамасы[2] орыс тілінің машиналық қоры дегеннің ғылыми түсінігі − көп тармақты әрі терең автоматтандырылған жүйе дегенді аңғартады. Бұл жүйе информациялық барлау міндеттерін де, сондай-ақ, тіл білімі саласындағы, кеңірек алғанда, жалпы филологиялық бағыттағы зерттеушілік міндеттерді де атқара алады. Орыс тілінің машиналық қорын жасаудың методологиялық негіздері ретінде екі тезис басшылыққа алынған: бірінші,− тіл жөніндегі деректердің қандайы болмасын лексикографиялық пішінде ұсынылады. Егер кейбір тіл жүйелерінің тарамдары немесе тұтас бір деңгейлері (мысалы, синтаксис) лексикография тұрғысынан сипаттауға әлі икемделіп жетпеген болса да, принциптік жағынан оны жүзеге асыру толық мүмкін деп саналады. Екінші, − лексикографиялық жағынан жүйеленген тілдік мәліметтердің барлығы да машинаға лайықталып, алгоритм тіліне аударуға келеді.

---

[1] Машинный фонд русского языка: идеи и суждения. М., 1986. С. 7-12.
[2] Бұл да сонда. 13−25-б.

Сонымен, тілдің машиналық қоры деген ұғымның түпкі төркіні лексикографиялық, яғни сөздік түріндегі жұмыстардың идеясына келіп тіреледі. Осы түсінікке сәйкес және басқа да ойларды қамти келе Ю.Н.Караулов машиналық қордағы тілдік информацияларды шоғырландыратын бірегей түйін немесе түп қазық жеке сөз деп есептейді. Өйткені, сөз тілдің ең табиғи бөлшегі болып табылады. Ол морфемалармен, сөз орамдарымен немесе сөйлемдермен салыстырғанда да ішкі-сыртқы қызметтері жағынан яғни морфемаға, буынға, әріпке, мағынаға талданатындығы, сондай-ақ, сөздерді өзгертетіндігі, жаңа сөз тудыратындығы, синтаксистік, грамматикалық, ассоциялық және концептуалды қатарлар жасайтындығы жағынан ерекшеленеді. Осыған орай, информация сақтағыштық қасиеттің негізінде мынадай бір кепілдеме болады: ол сөз − белгілі бір жүйедегі сөз қорын танытады, оның белгілі мөлшерде өзіндік сегменттері болады, олар грамматикалық, стилистикалық, фонетикалық, контекстік т.б. параметрлердің, демек, сол сөз жайындағы барлық мәліметтердің бүкіл жүйесін сәулелендіреді.

Мынадай екі жақты құбылысты қарастыра кетуге болады: тілдің машиналық қоры жасалып болғаннан кейін ол тек архив тәрізді қызмет ете ме, әлде машиналық тіл қоры тіл жөніндегі жаңа білімдер қосатын, жаңа мәліметтер тауып беретін зерттеу лабораториясы болмақ па? Бұл ретте айтарымыз: екі жақты құбылыстың соңғысы алғашқы қисынның барлық қасиеттерін иемденетіндігі себепті берер мүмкіншіліктері де молырақ. Демек, тіл қоры бізге белгілі тілдік заңдылықтар жөніндегі түсініктерді бойына жинақтаған тілдің синхронды және диахронды процестерін моделдей алатын әрекет үстінде болуы қажет. Сонда ғана машиналық қор түрлі деңгейдегі тіл бөлшектерін, жүйелерді негізге алып, қызмет ететін болады. Басқаша айтқанда тілдің машиналық қоры дегеніміз − барлық мәліметтерді жинақтаушы банк, бірақ ол «ақылды», мамандандырылған білім қоры болып табылады. Оны кітап ретінде пайдалануға болады, тиісті ұсыныстарын зерттеуші белгілі бір мақсаттарда іске асыра алады. Ғалымдардың айтуынша, машина қоры дегеніміз − ғылым адамының өз зерттеу объектісіне жаңаша тұрғыдан көз салудың мүмкіндігі болып табылады. Мұндай қор неғұрлым қомақтырақ болса, солғұрлым тіл құрылысының сырын тереңірек ашуға болады, сөйтіп зерттелетін объект жөніндегі түсініктердің шеңбері кеңиді, адамның білім өрісіндегі «ақтаңдақтардың» бедер-бейнесі айқындала береді.

Тілдің машиналық қоры адамдардың зерттеушілік қалып-қабілетін әлденеше есе арттырады, творчестволық қуат көздері ашылады, сөйтіп бұл жаңа мүмкіндіктер тілдің жүйелілік қасиеттерін жетілдіруге, яғни тіл жүйесін мұқият тануға жұмсалатын болады. Машиналық қор жасаудың программасы комплексті сипатта болуы тиіс, демек, ондай программа ең қажетті деген теориялық және экспериментті, яғни практикалық мәліметтерді бірлікте қамтуы қажет.

Машиналық қор жасайтын программа сонымен бірге белгілі бір нысананы көздеуі тиіс, өйткені, программаны жүзеге асырудың нәтижесінде жаңаша бір зерттеу объектісі, немесе материалдық жүйе, бұрын болмаған

жаңа аппараттар жасалып, тіл жүйелерінің барлық қасиеттері мен көріністері жан-жақты зерттелетін болады.

В.М. Андрющенко «Орыс тілінің машиналық қорын жасаудың концепсиясы мен архитектурасы» атты баяндамасында[3] тілдің машиналық қоры деген ұғымның ең жалпы түсінігі деп төмендегідей ұғындырады. Ол − тілді зерттеп, тексеретін автоматтандырылған комплексті жүйе болмақ. Оған түрлі лингвистикалық мәліметтер, сол мәліметтердің текстермен, картотекалармен, сөздіктермен, грамматикалық т.б. лингвистикалық білім көздерімен деректендірілген білімдер жинақталмақ. Осы білім көздерінің бәрі де программаға сәйкес қолданыс тауып жаңа тілдік объектілерді, яғни сөздіктерді, грамматикаларды, тіл дамуының түрліше процестерін құрастыруға қызмет етеді. Машина қоры жинақталған білім көздерінің бәрі де қорландырудың жаңа единицалары ретінде ғылыми мүмкіндіктерді молайта бермекші.

Тілдің машиналық қорын жасау үшін оны мынадай құрамдармен қорландыру керек: 1) машина қорының негізгі сөзтізбегі; 2) иллюстрациялық текстер қоры; 3) терминология қоры; 4) академиялық сөздік − грамматика қоры; 5) тіл процессорларының қоры; 6) лингво-статистикалық база; 7) лексикографиялық база; 8) лингвистикалық алгоритмдер мен программалар қоры.

Машиналық қордың осы айтылған жобасының ерекшелігі деп мынаны айтуға болады: бұл қор лингвистикалық мәліметтердің сала-салаға жүйеленген банк ретінде қызмет атқарады, яғни оның компопенттері дербес күйінде әр салада сақталады да, оларды тұтыну қажеттігіне сәйкес керек кезінде қайта жинақталады.

Бұл күндері жоғарыдағы ғылыми негіздерге сүйеніп СССР-дағы түрлі ұлттар тілінің машиналық қорын жасау да жоспарланып отыр. Бұл істің ұзын ырғасы он бес жылға созылмақ. Осы мерзім ішінде тіл зерттеуші барлық институттар мен ірі-ірі жоғары оқу орындарында лингвистикалық жұмыстарды автоматтандыру мен тіл мамандарына информациялық-анықтамалық қызмет көрсетудің қалыпты комплекстік жүйелері іске қосылуы тиіс. Тіл жөніндегі деректерді жинаудың, сақтаудың, талдаудың және салыстыра зерттеудің жаңа әдістеріне көшу, сондай-ақ лингвистикалық дерек көздерінің автоматтандырылған сөздіктер мен грамматикалар тәрізді жаңа түрлерін игеру ісі өмірге жанасымды және тиімді болу үшін бұл салада жалпыфилологиялық дәстүрлермен мәдениетке, тілді терең танып білуге, практикалық жаңа міндеттер мен тәжірибелерді мұқият ескеруге тиіспіз. Бұл ретте табиғи тілдің деректерін зерттеу жүйесіне негізделетін жаңа информациялық технологияны жетілдірудің бірден-бір шарты − адам мен машинаның қарым-қатысына келіп тірелетін интеллектуальды орта деп білеміз.

[3] Бұл да сонда. 26-44-б.

А.С. Гердтың «Орыс текстерінің типтері және орыс тілінің машиналық қорын құру» атты баяндамасында[4] тілдің машиналық қоры дегенді ол өзіне тән кішігірім жүйелер мен микросистемаларды қамтитын автоматтандырылған, аса үлкен макросистема деп түсіндіреді. Мұндай жүйелердің бірі сөйлеу фактілерін тіркеп отыруды нысанаға алатын тіл қорлары, атап айтқанда, текстер, сөздер, сөз формалары, сөз тіркестері мен сөйлемдер. Мұндай қорлар кішігірім жүйелердің тіркемелі типіне жатады.

Бұл жүйелердің мақсаттары мен міндеттері қандай болмақ? Мұндай жүйелерде тұтас текстер сақталады және тұтынушының сұранысы бойынша сол текстердің үзінділері, нақты адрестері және сөз формаларының статистикасы, сөз нұсқағыштардың (словоуказатели) толық тізімдері немесе солардың фрагменттері, сөздерге телулі цитаталар, жиілікті және кері алфавит сөздіктер беріледі. Осы арқылы сөздің әралуан тұлғаларының ең түпкі бірегей нұсқасы айқындалады. Болашақ машиналық қордың лексикасының сипаты түпнұсқа текстердің типіне байланысты болып келеді.

Бұл айтылғандардың кейбір мысалы ретінде яғни тілдің жүйесін танытатын сөйленіс жағдаяты деп мыналарды айтуға болады: көркем әдебиеттің тілі, газеттің тілі, ғылым мен техниканың тілі, ұйымдастыру-нұсқау документтерінің тілі, заң орындары мен сот мекемелерінің тілі, мектеп т. б. түрлі оқу орындарының тілі, жиналыстардың тілі, күнделікті әңгімелесудің тілі, транспорт, магазин, аптека, ателье, поликлиникалардағы сөйлеу тілі, жеке адамдардың күнделіктері мен қойын жазбаларының тілі т.б.

Әрине сөйленіс типтерінің бұл түрлері біркелкі емес, көпшілігінде жазбаша, ал енді біреулері ауызша жүзеге асады. Алайда жазба текстердің қорына даусыз енетін сөз үлгілері деп көркем әдебиеттің, ғылымның, техниканың, заңның, қаулы-қарардың, іс қағаздарының тілін айтамыз.

Қазіргі ғылыми-техникалық және іс қағаздары тексінің негізгі өзегі − терминдер, әйтсе де, қай тілдің болсын машиналық қорына ғылыми-техникалық текстерді түгел енгізу керек пе, жоқ па деген мәселе әлі де ойлануды қажет етеді. Жалпы сөйлеу тілінің аясындағы сөздерден емес, терминдердің тілдік қоршауы, яғни контексті өзгеше бір тілдік жүйе, басқаша айтқанда терминдер өрісі болып табылады. Осы аяда өмір сүретін термин үшін барлық ғылыми шығармаларды сөз қолданыс машықтары етене, ортақ болып келеді. Сырт қарағанда белгілі бір ғылым саласындағы терминдер өзінің аумағы жағынан бүкіл жасалатын қордан асып кететін тәрізді. Олай болғанда басқа ғылымдар саласындағы терминдерге машинаның жадынан үлес тимей қалуы да мүмкін. Сондықтан бұл мәселені терминологтардың алқасына салып, келелі кеңес жолымен ойланып-толғанып шешу керек. Себебі ғылым мен техника тілінің лексикасы дегеніміз тек термин сөздер ғана емес, бұл ұғымға ең алдымен бүкіл терминдік тіркесімдер, символдар, теңгерме элементтер немесе идентификаторлар да кіреді. Бұлардың саны күн санап өсе бермек, өйткені бұл қызметті атқаратындар кәсіби етістіктер, шартты символика т. б. болады да олардың әрқайсысы белгілі бір ғылыми

[4] Бұл да сонда. 67−75-б.

ұғымдар мен концепцияларды танытады. Машиналық қорға текстер енгізуде қиынырақ тиетін мәселелер диспуттар, сот процестері, жиналыстар, күнделікті әңгімелесулер, реплика тастау, айтыс, телефон арқылы сөйлесу, радиомен хабарласу кезіндегі ауызекі сөйлесу тілінің үлгілерін жазып алу болып табылады.

Текстердің әр дәуірдегі белгілерін ажыратып, тануда және оларды сол дәуірлер бойынша жазып алу аса маңызы жұмыстардың бірі. Текстердің хронологиялық жігін айырып тану мақсатында диалектілерді машиналық қорға қосудың ойланатын жақтары жоқ емес. Диалектілер синхронды түрде өмір сүреді. Бірақ өздерінің мәні жағынан олар әрі диахронды болып келеді. Диалект сөздер кез-келген ұлт тілінен әсіресе әдеби тілден ескі, әріден келе жатқан тіл құбылыстары. Ал, қазіргі тіліміздегі диалектілердің лексикасы біртектес емес, әркелкі. Олар бір жағынан ертедегі сөздер мен қолданыстардың ұлтқа, жазуға дейінгі үлгілері болып келсе, екінші жағынан басқа тілдерден енген сөздерді тіл ішінде кейіннен пайда болған, бірақ оқшау, томаға-тұйық қолданыста болып, таралу өрісі шектеліп қалған сөздер болуы да ықтимал. Халық тілі сол халықтың жанды қазынасы болғандықтан соның бүкіл болмысын барлық ерекшеліктерімен қоса қамту машиналық қордың міндеті.

Диалектілер секілді әрі синхрониялық әрі диахрониялық сипатта күн кешетін тіл − фольклордың тілі және жер-судың тілі немесе топонимика. Осы аталған тіл үлгілерін жинастырып, машиналық қор жасағанымызда, оның бұл аталғандардан басқа да салалары мен арналары қамтылары сөзсіз. Қысқасы машиналық қор жасау мақсатымен байланысты бүгінгі таңда күн тәртібіне қойылатын мәселелер мыналар деп білеміз: тіл маманына ең алдымен қажет болатын мәліметтердің сипаты. Керек болған белгілі бір сөзді қандай тіл көздерінен, текстерден тіл айырымдарынан іздестіру қажет және сол сөз тиісті тексте қамтылған ба, жоқпа, графикалық және фонетикалық пішіндерін қоса есептегенде сөздің мүмкін болған барлық түрлері мен тұлғалары айқындалып болды ма, белгілі бір сөзге қатысты контекстердің жалпы аумағы, сипаты анықталды ма, белгілі бір сөз осы семантикалық және грамматикалық сипатта өткен дәуірлердегі, көне ғасырдағы текстерде қолданылған ба, егер қолданылса, қай замандардан бергі сөздіктерден көрініс тапқан.

1988 жылдың май айында Москвада Совет тюркологтары комитетінің XIV Пленумы болып өтті. Пленум мәжілісінің күн тәртібінде: «Түркі тілдерінің машиналық қорын жасау» проблемалары талқыланды. Онда негізгі баяндамашылар: Р.Г. Пиотровский, А.М. Щербак, В.Г. Гузев өз ой-пікірлерін ортаға салып, түркі тілдерінің машиналық қорын *(бұдан әрі – ТТМҚ)* жасаудың қажеттілігімен қатар қиыншылықтарын да атап көрсетті[5].

ТТМҚ-ын жасаудың болашақта екі жақты қажеттілігі (ұтымдылығы) болатындығы ескерілді. Біріншіден, салыстыру-салғастыру (сравнительно-

[5] *Гузев В.Г., Питровский Р.Г., Щербак .А.М.* О создании машинного фонда тюркских языков // Советская тюркология. 1988. № 2. С. 98-101.

сопоставительный) тұрғысынан қарастырғанда лексикографиялық, грамматикалық, стилистика-статистикалық, тарихи-этимологиялық мәліметтерді жинау және реттеу (тез іздеп табу); екіншіден, тұтынушыларға (тілші-зерттеушіге) бір тілдің тұрақтты единицаларының «туу» (порождение) ережелерінін көмегі арқасында қажетті тілдік элементтерді әралуан текстен автоматты түрде бөліп алу мүмкіншілігіне ие болу.

Түркі тілдеріне ортақ машиналық қордың маңызы сөз бола келе, сол қорды жасаушылар, яғни тілшілер мен инженер-программистер теориялық жағынан маңызы зор екі түрлі әдістің біреуін: «жоғарыдан төменге» (нисходящий) немесе «төменнен жоғарыға» (восходящий) қарай зерттеудің жолын күн ілгері таңдап алу қажеттігі сөз болды.

Қомақты жүйелерді құрастыру кезінде (ТТМҚ-ын құрастыру да соған жатады) «төменнен жоғарыға» қарай тексеру принципінің елеулі қиыншылықтарға әкеліп соғатыны нақты ескертілді. Себебі бұл принцип бойынша автоматты жүйе құрастырғанда бөлек-бөлек жағдайларға (есептерге) негізделген қарапайым модульдер жасалып, олар бірте-бірте бірігіп, іріленіп, күрделі блоктарға айналып, нәтижесінде кішігірім жүйелердің орнына қомақты біртұтас жүйе құру мақсаты көзделеді. Алайда мұндай бөлекше модульдер өз алдына жеке тұрып жұмыс істегенмен біртұтас біріккен жүйеге (демек бүкіл түркі тілдерінің машиналық қорын жасау жүйесіне) айналғанда алға қойған мақсатты шешетіні, шешпейтіні бимәлім. Көп жағдайда мақсат орындала да бермейді. Басқаша айтсақ, жеке тұрып лексика-семантикалық, этимологиялық, морфологиялық, тіпті, синтаксистік есептерді шешетін программалардың жиынтығын және ЭВМ жадына жазылған неше түрлі лексикографиялық материалдардың тұтастығын тілдің машиналық қоры осы деп есептеуге болмайды.

Осы айтылғандарды қорыта келе, авторлар түркі тілдерінің машиналық қорын жасау деген ұғымды екі түрлі концепция (стратегия) тұрғысынан қарау қажеттігін ұсынды. Бірінші концепция түркі тілдерінің ғылыми машиналық қорын жасау жайында болса, екінші концепция бойынша түркі тілдерінің көп аспектілі машиналық қорын жасау міндеті көзделеді.

Мұндай концепцияларға бөлу мына көзқарастан туады. Біздегі ғылыми-зерттеу орталықтарының нақты жағдаяттарын ескеретін болсақ, тіл мамандарының өзіне бекітіліп берілген ЭВМ орталықтарының жоқтығы, қажетті маман инженер-программистер мен тілшілердің жетіспеуі т.б. салдарынан таяудағы 5-7 жыл шамасында түркі тілдерінің көпаспектілі машиналық қорын жасау мүмкіншілігі екіталай. Сондықтан да жұмысты түркі тілдерінің шағын көлемді ғылыми машиналық қорын жасаудан бастау жөн болады.

Осы тұрғыдан қарағанда, ең алғаш түркі тіліндегі сөздердің грамматикалық формаларының жасалу (порождения) процестерін зерттеумен шектелген жөн. Толығырақ айтқанда ғылыми машиналық қор түркі тілдеріндегі бір буынды сөздердің әртүрлі құрылымдық-фонетикалық үлгілерінің толық жиынтығын қамтуы тиіс және ол дыбыстардың тіркесуі мен қисындасуы жағынан ата түркілік күйден шет шықпағаны жөн. Бұл

жиынтық әрі тұрақты, әрі өзгермейтін болуы шарт. Бір буынды сөздердің алғашқы қалпын анықтау фонетиканы салыстырмалы анализ әдісімен зерттеуді қажет ететіні түсінікті. Түркологтердің компаративистік зерттеулері негізінде алған нәтижелерінің өзі ата түркілік күйдегі сөз тұлғасының біршама тиянақты моделін құруға мүмкіндік береді. Ал, модельді жетілдіре түсу жолдарын табу үшін жаңа зерттеулердің қажет екендігі түсінікті.

Екінші концепция бойынша түркі тілдерінің көпаспектілі машиналық қорын жасау қолданбалы ілімдердің мұқтажына сай жүзеге асырылуы тиіс. Мысалы, оны баспа, радио, телевизия, театр қызметкерлерінің, мұғалімдердің түркі тілдерін пайдалану қажеттігіне орайластыруға әбден болады. Мұндай мұқтаждықты өтеу үшін түркі тілдерінің көпаспектілі машиналық қоры әрі жалпы-түркілік тілдің жиынын, әрі жекелеген нақты бір тілдердің жиынын модельдеу арқылы жасалады.

Төменде осы айтылған машиналық қордың түрліше салалары (блоктары) жалпылама сипатталмақ.

1. Ең алғаш құрастыруды қажет ететін нәрсе – лексикалық блок. Ол орфографиялық, орфоэпиялық, түсіндірме, терминологиялық қолданыс жиілігі және аударма (орысша-түркіше, түркіше-орысша) типіндегі сөздіктерді қамтитын көпшілік қолды лексикографиялық жүйеден тұруы тиіс. Бұл ретте жалпытүркі лексиконын жасау (құрастыру) ісінің өте-мөте қиындыққа түсетінін ескеру қажеттігі сөзсіз.

2. Грамматикалық блок – лексикалық және грамматикалық морфемалардың тізбелерінің жиынтығынан, сондай-ақ синтаксистік схема-модельдерді қамтитын тілдік материалдардан тұрады. Әрбір жеке тіл бойынша морфемалармен синтаксистік схемалардың жиынтығы қосыла келе жалпытүркілік тілдің машиналандырылған грамматикалық тезаурусын жасауға мүмкіндік береді. Оның түркі тілдерінің ғылыми-машиналық қоры ретінде де кәдеге жарайтындығы ескерілмек. Бұған типті тезаурустың түп-тірегі ретінде аффикстер мен грамматикалық көрсеткіштердің жиынтығы алынғаны жөн.

3. Түркі тілдерінің машиналық қорындағы ең бір шағын жүйе − фонологиялық және морфонологиялық блоктар.

Фонологиялық блок әр тілдің фонемалық құрамын (состав), әр фонемаға тән бүкіл дифференциалдық (айырым) белгілердің түсіндірмелерін қорландыратын жиынтықтан тұруы қажет.

Морфонологиялық блок төмендегідей информациялық мәліметтерді қорландырып сақтайды: 1) тілдік единицалардың фонемдік құрылысының типтері: негіздер, форманттар, сөздердің жекелеген кластарына (мысалы, бейнелеуіш сөздерге) тән фонемдік құрылыстың ерекшеліктері жөніндегі деректер; 2) белгілі бір тіл үшін міндетті сингармонизм заңдары, сондай-ақ дауыссыз дыбыстардың ассимиляциясы, қатаңдауы мен ұяңдауы жайындағы мәліметтер; 3) түбір-негіздер мен сөзформаларға түсетін екпін туралы мәліметтер т.б.

Әралуан түркі тілдеріне қатысты морфонологиялық суреттемелерді салыстыру нәтижесінде жалпы түркілік морфонологиялық белгілердің терең сырларын ашуға, сондай-ақ жекелеген тілдердің өзіндік қасиеттерін тануға қажетті мүмкіндіктер туады.

Міне осы аталған мәселелердің бәрін жан-жақты қамти отыра толық та тұжырымды жауап алуымыз үшін электронды есептегіш техниканы пайдалана отырып, тілдің машиналық қорын жасау − лингвистикалық ғылым саласындағы қайта құрудың теориялық әрі практикалық міндеттерді жаңаша ойлап, жаңаша шешудің бүгінгі таңдағы ең тиімді жолы болып табылады.

ТЕКСТ ЛИНГВИСТИКАСЫ ЖАЙЫНДА

Қаз. ССР ҒА Тіл білімі институты. «Лингвистикалық жұмыстарды автоматтандыру және тіл статистикасы» тобы қызметкерлерінің еңбектері: «Қазақ тексінің статистикасы». II шығарылуы. Алматы, «Ғылым», 1990, 118-125-бб.

Қазіргі кезде әр түрлі тілдерді жан-жақты зерттеудің нәтижесінде өмір талаптарына сәйкес бірнеше тілдік жүйелер іс жүзінде пайдаланылып жүр. Солардың басты-бастыларына мына төмендегілер жатады:

1. Бір тілден екінші тілге аудару жұмысын автоматтандыру жүйесі.
2. Жазба мәліметтерінен автоматты түрде аннотация және рефераттарды (мәліметтердің қысқаша мазмұны) ықшамдап алу жүйесі.
3. Хабарлама (информация) жұмыстарын автоматтандыру жүйесі (қажетті хабарды автоматты түрде іздеу және оны сақтау мәселелері) [1, 5-б].

Осы аталған тілдік жүйелерді пайдаланудың ойдағыдай шешілуі − сол табиғи тілдің функционалды моделінің бірден-бір дұрыс құрастырылуына тығыз байланысты [2–4].

Бұл жүйелер қай тілде іске асырылса да, олар өздері іштей белгілі бір принципке бағынады. Олар атап айтқанда сөздердің мағынасын ажырату арқылы сөйлемнің мазмұнын ашу, одан соң сөйлемдер жиынтығының мағыналары негізінде текстің мазмұнын айқындауға саяды. Әрине, бұдан сөйлем мазмұны сөздер мағыналарының қосындысы деп, ал текст мазмұны сөйлемдер мазмұнының жиынтығы деген бірбеткей тұжырым тумауы керек. Біз тек текстердің мазмұнын автоматты түрде ашу өте күрделі және арнайы зерттеулерді қажет ететіндігін ғана атап айтпақпыз.

Текстер жайлы бірер сөз. Көптеген ғылым салаларындағы кейбір ұғымдардың тиянақты анықтамасын дәлме-дәл берудің қиындығы сияқты, «текст» ұғымына да тұрақты бір анықтама берілмей келеді. Мұның дәлелі ретінде осыдан он-он бес жылдар бұрын бұл ұғымның ешбір анықтамасыз-ақ қабылданып келгенін айтсақ та жеткілікті тәрізді [1, 9-б.]. Қазірдің өзінде тілші-мамандардың «текст» ұғымын сипаттайтын анықтамалары әр қилы. Мысалы, «текст дегеніміз − «ұзын сөйлем», «текст − ол бір-бірімен

мағыналық қатынаста болатын, кемінде екі сөйлемнен тұратын, сөйлемдер жиынтығы», − деген сияқты анықтамалар беріліп жүр. Басқа да, күрделіленілген неше түрлі информациялық, логикалық және лингвистикалық ұғымдарды пайдаланып айтылып жүрген анықтамаларды да кездестіруге болады. Мысалы, семиотикалық тұрғыдан қарастырғанда «текст − тұтас бір мағыналы кез келген белгілер құрылымы» (знаковая структура), немесе «текст − семантикалық жағынан ұйымдасқан белгілер жүйесі» − деген анықтамалармен түсіндірілген [5, 154-б.]. Ал, логика-лингвистикалық тұрғыдан құрылған анықтамаға тоқтасақ − «текстің» басты белгісі ретінде сызықтық тізбе құрайтын сөйлемдерден тұратындығы, сол жақтық және оң жақтық шекарасы болатындығы, одан соң салыстырмалы толықтық пен байланыстық шарттарының орындалуы тәрізді ерекшеліктер негізге алынады [6, 172-б.].

Ғалымдардың «текст» ұғымы бойынша келтіріп жүрген әр түрлі анықтамаларын қорыта келе, текстің мынадай негізгі қасиеттерін атап кетуге болады. Ол − текстің бір мазмұнды болуы және оның единицаларының өзара грамматикалық, семантикалық әрі логикалық байланысқа түсуі.

Текстерді − байланысты не байланыссыз деп ажыратудың негіздеріне сүйене отыра оның сыр-сипатын ашуға арналған анализ-зерттеулер соңғы жылдарда синтез-зерттеулерге қарағанда көбірек жүргізіліп жүргенін аңғаруға болады.

Текстердің құрылу заңдылықтарын анықтау, әсіресе олардың ішінен ең негізгі деген единицасын айыра білу − өмірдегі оқыту процесінде тәжірибелік мәні зор. Дәлірек айтқанда, оқушылардың аз уақыттың ішінде көлемді текстердің мағынасын дұрыс меңгере білу, жаңа программаға сай оқулықтармен оқу құралдарын жазу үстінде аса қажет деуге болады. Сондықтан да әр түрлі текстердің ішкі заңдылықтарын егжей-тегжейіне дейін зерттеу және оларды аша түсу бүгінгі күннің алға қойып отырған талабы десек те артық болмас.

Текст құрылымын зерттеудің қажеттігі, ең әуелі оның негізгі единицаларын анықтап, айыру үшін текстерді сегменттеу ережелерін тағайындап алып, ірі-ірі текстік единицаларға топтау жағын қарастырсақ жөн. Зерттеуші ғалымдардың ұсыныстары бойынша осындай текст құрайтын ірі единицаларға: сөздер, сөз тіркестері, сөйлемдер, абзацтар, параграфтар, бөлімдер, тараулар, томдарды жатқызуға болады [1, 12-б.].

Өмірдің жеке-жеке объектілердің арасындағы кейбір қарым-қатынастарды айқындау барысында жалаң сөздерге қарағанда сөз тіркестерінің рөлі аса зор екендігі белгілі. Алайда сөз тіркестері әр уақытта сөйлемдер сияқты аяқталған бір ойды білдіретін хабарлама бола бермейтіндігі белгілі. Өмірдің объектілері сапалық және сандық сипаттамаларымен қосарлана отырып, белгілі бір жағдайларға қатысты объектілермен қарым-қатынаста болып отырады. Осындай ситуация құрайтын ерекшеліктер тексте атрибутты, объективті, предикативті, адвербиальды сөз тіркестері түрінде болады. Мысалы, адам баласының жас кезінен-ақ оңай және тұрақты түрде есінде сақтайтын сөз тіркестері болады:

ақ қар, көк аспан, ана сүті, құс ұшады т. б. көптеген сөз тіркестері [7, 27-б.]. Міне осындай сөз тіркестері сияқты текст құрылымының элементтері ретінде зат пен оның қасиетін суреттейтін әр түрлі тілдік тәсілдер, соның ішінде, көбінесе, тексте қолдану жиілігі жағынан молырақ кездесетін сөз тіркестерінің түрлері алынып жүр.

Сөз тіркестері өздерінің жасалуы жағынан болсын не құрылымы тұрғысынан болсын құбылмалы шама бола тұрса да, әр авторға қатысты өзіндік айырым қасиеттері де болмай қоймайды. Көпшілік тілдерге тән басқа да ортақ заңдылықтардың кездесе беретіні тәрізді, сөз тіркестерінің, қатарындағы ең жиі кездесетіндері екі элементтен тұратын тіркестер екені де жалпы тілге тән заңдылықтар сияқты. Тіркестердің мұндай түрлері, көбінесе, зат пен оның қимыл-әрекетінің арасындағы қарым-қатынастарын айқындайды.

Тіл жүйесінде осы күнге дейін толық шешімі табылмай жүрген мәселелердің ұзын-ырғасы аз емес. Солардың бірі − сөйлемнің жалпы тіл жүйесіндегі алатын орны, екіншісі сөйлемнің семантикалық және синтаксистік жасалу принциптері. Сол сияқты, сөйлемдегі текстік негізгі семантикалық компоненттерінін қатарына жатқызу не жатқызбау мәселесі де әлі өз шешімін таппай келе жатқан жайдың бірі [1, 16-б.].

Текстегі абзац мәселесінің де тілші ғалымдар арасында әлі де көп талас тудырып жүргені мәлім. Бүгінге дейін абзацты текстің қай единицасына жатқызамыз деген сұраққа әр түрлі пікірлер айтылып келді. Оқымыстылардың кейбірі оны композициялық десе, екінші біреулері − логикалық, не семантикалық единица деген топқа қосып жүр. Енді бір зерттеушілердің айтуынша абзац − бір орынға жай ғана топталған сөйлемдердің жиынтығы емес, ол автордың өз ойын оқушының саналы түрде қабылдауына, оны меңгерудің амалдарын жеңілдетудің мақсатына лайықтап құрған сөйлеу практикасында бар текстік единица дейді.

Абзац мәселесін мазмұн тұрғысынан қарағанда көпшілікке мәлім мынадай фактіні атай кетуге болады. Әр текстің тақырыбы іштей бөлек-бөлек тақырыпшаларға, субтемаға және микротемаларға таралады [8, 27-б.]. Тексте автордың ойы шағын-шағын үзік хабарламалар арқылы беріледі. Ал, осы хабарламалардың бірігіп қосылуынан қомақты хабар пайда болады. Осылайша біріккен хабарламалардың ең шағыны микротема деп аталады да, ал барлық текстің тақырыбын белгілі қатынастық байланыста тұратын бірнеше микротемалардың қосындысы деп есептейді.

Абзацқа енетін әр түрлі текстерден алынған сөйлемдерді байланыссыз бірнеше абзацтардан алынған сөйлемдер тізбегімен салыстыра қарастырғанда төмендегідей ерекшеліктерді байқауға болады: а) абзацқа қатынасты сөйлемдер бір микротеманың не болмаса бір-бірімен тығыз байланысты тұрған бірнеше микротеманың мазмұнын анықтайды; ә) бір абзацтың ішіндегі сөйлемдер арнаулы лексикалық амал-тәсілдер арқылы өзара байланыста болуы керек; б) абзац ішіндегі сөйлемдердің бір-бірімен белгілі синтаксистік байланыста болуы да шарт [9, 14-б.].

Сөйлеудің (речь) функционалды-мағыналық түрлеріне қарай текстегі абзацтарды төмендегіше топтастыруға болады: 1) хабарлау абзацтары; 2) сипаттау абзацтары; 3) пікір тудыру абзацтары; 4) сұрақ мәнді абзацтары; 5) ниет, ойды білдіретін абзацтар; 6) аралас абзацтар [10, 20-б.]. Құрылымдық көзқарастан алып қарағанда абзацтың лексикалық тұрғыдан толыға түсуі, синтаксистік және грамматикалық жағынан түрлене түсуі сол абзацтың текстің жалпы құрылымындағы тұрған орнына да байланысты болады. Осы себептен абзацтар: кіріспе (бастапқы), ішкі және қорытынды абзацтар болып үшке саналады.

Шет ел тіл мамандарының да, отандас тілші ғалымдарымыздың да лингвистикалық зерттеулерінің жеке сөйлем көлемінен тысқары шыға қоймағандығын көбірек байқауға болады. Тек соңғы жылдардағы «текст теориясы», «текст лингвистикасы» сияқты атпен шығып жүрген ғылыми ізденістер текстің пайда болу (туу) мәселелерін сөйлем көлемімен ғана шектелмей текст шеңберінде сөз етіп жүр [11, 10-б.].

Жедел дами бастаған текст теориясы көптеген күрделі проблемаларды кеңінен қарастыруға жол ашты. Бұл проблемаларға, қысқаша айтқанда, текстің негізгі единицаларының басын ашу, саралау және олардың өзара байланысының сыр-сипаты мен шарттарын белгілеу және де текстің мазмұнын құрайтын тақырып пен теманың ролін ашу мәселелері жатады.

Жоғарыда сөз болғандай тілші психологтерден кейін көп ұзамай-ақ тілшілердің өздері де текст бойындағы жалпы тақырыптың шағын тақырыпшаларға (подтема, субподтема, микротема) ажыратылатындығын қолдай бастады [12, 25-31-б.].

Зерттеушілер текстің пайда болу (туу) процесінің алгоритмін бір арнаулы жүйелі тәртіп (процедура, диспозиция) ретінде қарастырғанды мақұлдайды. Осы жағдайда ғана сөйлем құрудың әр түрлі жағдайлары мен жасалу шарттары ескерілуі мүмкін. Әрине, жоғарыда аталған алгоритм құрудың жүйелі тәртібін қатаң, бірізді операциялар деп түсінбей, оны таңдау мүмкіншілігі мол ықтималдық ауысулардың негізінде іске асатындығын ескеру қажет [10, 121-б.].

Байланысты текстер жайлы теория, әсіресе сөйлемдердің өзара тіркесу заңдарының синтаксистік тұрғыдан баяндалған нақты ережелері көрсетілген зерттеулер әлі де болса жетіспейді. Жарық көрген еңбектерде текстің «туу» мәселесіне келгенде ғалымдар төмендегідей қорытынды жасайтынын байқауға болады. Текст, тек қана сөйлемдер мағынасымен шектелмейтін, одан да ірілеу текст бөлшектерінің белгілі реттілікті сақтайтын фразадан жоғары (сверхфраза) бөліктермен мағына түзеді [13].

Автордың белгілі бір жайды тұжырымдап айтып жеткізу қажеттігі кезіндегі сөздерді орын-орынымен, рет-ретімен таңдай білу мүмкіншілігінің сырын ашып айту жолдары да оңайға түспейді. Әдетте мамандар сөздер жиынтығын үлкен екі топқа бөледі. Олардың біріншісін мағыналық сөздер тобына жатқызса, ал екіншісін − мәнерлеуіш сөздерге жатқызады [14, 44-б.]. Бірінші топтағы мағыналық сөздердің қалайша сөйлемдердегі синтаксистік құрылысты (конструкция) толықтыра түсу жағын (наполняемость) өз алдына

жеке зерттей келе, оқымыстылар ол жағдайдың арнайы логикалық программаның негізінде іске асатындығын және сөйлемде қатынасатын тіл элементтерінің жиілігі мен өзара байланыс күшінің ролін атап айтады.

Автордың текст «жасау» (тудыру) тәжірибесін семантикалық және грамматикалық тұрғыдан зерттеп, сол мақсатқа қалайша тез жетуге болады деген заңды сұрақ тууы мүмкін. Әрине «тез» деген сөз бұл жерде орынды болмауы да мүмкін, бірақ біз әр түрлі мүмкіншіліктермен салыстырғанда әрі тез, әрі дәл түсетін жолды таңдауымыз қажет. Тәжірибе нәтижелеріне қарағанда аталған мақсатқа сатылап жететініміз байқалды [15, 22-26-б.]. Мысалы, сөйлемдегі сөзқолданыстардың атқаратын синтаксистік әрі морфологиялық қызметіне қарай ірі-ірі топтарға бөлуге болар еді. Керекті сөзді таңдап алу үшін оның сөйлемдегі атқаратын функциясын білген жөн. Мысалы, қажетті таңдайтын сөзіміздің сөйлем мүшесі ретінде атқаратын қызметі бастауыш не толықтауыш болу қажет десек, онда ол сөздерді бір зат есімге, не есімдікке жататын сөздер тобынан іздестіргеніміз дұрыс болады. Сол сияқты, кезекті сөздің сөйлемдегі атқаратын қызметі баяндауыш ролінде болса, ондай сөздерді етістік болатын сөздер тобынан, ал анықтауыш сөздерді − сын есім, есімше не болмаса ретті сан есім сөздер арасынан іздеген орынды болады.

Бәрімізге мәлім жайт − ол зат есім, етістік, сын есім және тағы басқа сөз таптарына қатысты сөздердің сан жағынан алғанда өте-мөте мол кездесетіндігі. Егер де әр сөздер тобындағы сөздерді тағы да кішілеу топтарға бөлуге мүмкіндік жасайтын қосымша белгілер табылса, сөз таңдау процесі тездетілген болар еді. Мысалы, орыс тілі үшін зат есімнің қосымша белгілері ретінде оны жанды және жансыз деп екі топқа бөліп алу мүмкіншілігі бар. Енді осы топтарды әрі қарайда шағын бөлшектерге бөлуге мүмкіндік беретін басқа да табиғи құбылысының белгілерін іздестіруге болатыны сөзсіз. Осындай семантикалық критерийлердің көмегімен етістік, есімше және т. б. сөздердің де іздестіру зонасын (өрісін) шағындатуға әбден болар еді. Ал, сөздерді көлемі кішірейтілген сөз топтарынан таңдап алу мәселесі текстердің стильдік құрамында болатын ықтималдық теориясының мүмкіншіліктеріне негізделеді. Әрине, осындай жолмен таңдалып алынған сөз басқа да факторлармен тексерілгені жөн.

Сонымен, ғалымдар адам баласының текст құру процесінің моделіне сүйене отырып, «қолдан» автоматты түрде текст «тудыру» есебінің алгоритмін құруды сөз еткенде мына жағдайлардың беті ашылғанын жөн көреді: 1) формальды түрде текстің жалпы тақырыбының берілуі; 2) текстің негізгі семантикалық-синтаксистік единицаларының бөлініп алынуы; 3) осы единицаларды байланысты текст түріне біріктіретін ықтималды-детерминделінген ережелерінің берілуі [1, 69-б.]. Осы мақсаттың орындалуы әр түрлі құрылымдағы текстер түрлері мен типтерін арнайы тәсілдермен терең зерттелуіне көптен-көп байланысты. Бұл жағдайларда, зерттеуге алынған текстердің саны қанағаттандырарлықтай болуы да шарт.

Біз осы ұсынып отырған кіріспе мақаламызда жалпы текст теориясындағы кейбір алғы шарттарды қазақ оқырмандарына арнап сөз етуді өзімізге мақсат деп білдік.

Қазіргі кездегі партия мен үкіметтің ұлттар тілінің дамуына зор мән беруі тілші ғалымдардың өз ісіне жаңаша мағынамен қарауын қажет етеді. Бұлай айтуымыздың бір себебі − техникалық прогрестің жаңа жетістіктерін пайдалана отырып, зерттеу мен және оның нәтижелерін дұрыс тұтыну қажеттігі болады деп есептейміз. Міне осы айтылғандарға байланысты қазақ текстерінің өзіне тән жасалу (туу) моделін құру қазіргі заманға сай үлкен мақсаттардың бірі болмақ.

Болашақта көптеген іс қағаздарының, неше түрлі хабарлардың және де ғылыми-техникалық әдебиеттердің қазақ тілінде басылып шығуы ықтимал. Олай болса қазақ тіліндегі текстерді автоматты түрде − компьютер көмегімен қажеттілікке пайдалану мақсаты тіл зерттеу тәжірибесіндегі жаңа сала − текст теориясымен тығыз шұғылдануды алдын ала бастаған өте орынды деп санаймыз.

Қорытынды ретінде, осы аталған мақсатты іске асыру қажеттігімен қазақ тексін абзац шеңберінде семантикалық-синтаксистік тұрғыдан зерттеу тәжірибесінің бастапқы қадамдары жайлы қысқаша тоқталмақпыз.

Қазақ тексінің құрылымын зерттеуді біз үш түрлі стильді қамтитын текстердің негізінде қарастырмақпыз. Ол − публицистикалық стиль, көркем әдебиет және ғылыми-техникалық стильдерді қамтитын текстер.

Публицистикалық стиль ретінде «Қазақстан мұғалімі» газетіндегі 1986 жылдың қаңтар айы мен 1987 жылдың қаңтар айы аралығындағы мектеп реформасына байланысты: «Совет мектебі жаңа кезеңде», «Алты жасарлар мектепке барады» атты рубрикамен басылып шыққан 35 мақала қарастырылады. Мақаланың авторлары − Қазақстандағы мектеп мұғалімдері, арнаулы газет тілшілері, мұғалімдер білімін жетілдіру институттарының қызметкерлері және тілші ғалымдар. Аталған 35 мақала 503 текстік абзацтан тұрады.

Көркем әдебиеттерге жататын текстер қазақтың белгілі жазушыларының шығармаларынан алынған келесі әңгімелер:

1) С. Сейфуллин. Бандыны қуған Хамит, 102 абзац;
2) Б. Майлин. Қанды кек, 98 абзац;
3) М. Әуезов. Көк серек, 97 абзац;
4) С. Мұқанов. Әсия, 47 абзац;
5) Ғ. Мүсірепов. Жапон балладасы, 92 абзац;
6) С. Ерубаев. Бахыт, 53 абзац.

Осы аталған алты әңгімеде барлығы 500 абзацтан тұратын 20 шағын текстер кездеседі.

Ғылыми-техникалық стильге жататын текст ретінде екі монографияның ішіндегі текстер алынды [16].

Екі монография бойынша алынған текст ұзындығы − 500 абзац.

Сонымен, үш стильге қатысты текстер бойынша 1500 абзац құрылымы зерттелмек.

Тәжірибенің алғашқы қадамдары төмендегі жұмыстарды қамтиды:

1. Аталған текстердегі сөзқолданыстарды тиісті сөз табына жатқызып белгілеу.

2. Сөз формаларды реестрлік сөзге келтіріп сегменттеу.

3. Әр текстегі мынадай төрт түрлі синоним сөздерді айырып белгілеу: лексикалық синонимдер, контекстік синонимдер, есімдіктік синонимдер, ассоциативті синонимдер.

4. Бөлек стиль бойынша әр абзацтың алфавитті-жиілік сөздіктерін алу.

5. Стильдер бойынша әр тақырыпшаларға қатысты сөздердің жалпы жиілігі мен абзацтар санына тиісті жиілігі бөлініп көрсетілген − үлестірімді алфавитті-жиілік сөздікті алу.

6. Үлестірімді сөздік бойынша сөздер жиілігінің кему ретімен түзілген − жиілікті сөздік алу.

7. Негізгі сөздердің потенциалды сөздігін алу.

Электронды есептегіш машина жәрдемімен алынатын осы сөздіктердің материалдарын пайдаланудың негізгі мақсаты – әр текстің, әр стилдің негізгі элементтерін, яғни текстің негізгі сөздерін (опорные слова) айырып алу қажеттілігі.

Осы этапта алынатын сөздіктердің материалдарының сапалылығы кейінгі этаптардағы тәжірибелер нәтижелерінің құндылығын арттыратыны сөзсіз.

Сонымен текстердің негізгі сипаттамаларын дұрыс анықтау әрі қарайғы абзац шеңберіндегі семантикалық және сиптаксистік зерттеулердің бірден-бір бастамасы болмақ.

Әдебиеттер

1. *Зубов А. В.* Проблемы порождения текста. Ч. 1. Теория и алгоритмы. Минск, 1987. С. 77.

2. *Виноград Т.* Программа, понимающая естественный язык. М., 1976.

3. *Марчук Ю. Н.* Проблемы машинного перевода. М., 1983.

4. *Невежина З. В.* Структурно-семантическая организация сверхфразового единства в современном английском языке. Киев, 1971.

5. Russion literature − the Hague Paris, 1972. P. 2.

8. *Hartmann P. Zur* antropoligiachen Fundiering der Sprache // Sumbolae Linguistical in honorem G.Kurilovich Warsszawa. 1965.

7. *Мер-Минасова С. Г.* Словосочетание в научно-лингвистическом и дидактическом аспекте. М., 1981.

8. *Марчук Ю. Н.* Проблемы лингвистического анализа связного текста (Межфразовый уровень). Таллинн, 1975.

9. *Аграскина С. Н.* Структура абзаца в художественной прозе: (на материале романов Э. Хэмингуэя «И восходит солнце» и «Прощай оружие»). Л., 1972.

10. *Кацнелсон С. Д.* Типология языка и речевое мышление. Л., 1972.

11. *Аспекты* общей и частной лингвистической теории текста. М., 1982. С. 192.
12. *Новиков А. И.* Семантика текста и ее формализация. М., 1983.
13. *Гальперин И. Р.* Текст как объект лингвистических исследований. М., 1981.
14. *Проблемы* текстуальной лингвистики. Киев: Изд-во КГУ., 1983.
15. *Пиотровский Р. Г.* Инженерная лингвистика и теория языка. Л., 1979.
16. *Қалыбеков Т.* Экология және ашық кен. Алматы, 1988: *Көшербаев К.* Минералдар әлемінде. Алматы, 1986.

ФОРМАЛИЗАЦИЯ ПРОЦЕДУРЫ ПОСТРОЕНИЯ ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА

Совет по автоматизации, вычислительной технике
и научному приборостроению при президиуме
АН КАЗССР. Институт математики и механики.
Сб.: «Научные приборы и автоматизация научных
исследований». Алма-Ата. «Ғылым», 1992. – С. 190-192.
(Соав.: А.В.Зубов – Минск. Гос.Пед.Инст. Иностр-х яз-в)

НТР, особенно интенсивно развивающаяся со второй половины XX века, вплотную поставила перед учеными задачи изучения систем высокой степени сложности. К ним относятся и живые организмы, и общество, и язык. Исследование этих систем требует разработки новых теорий, учитывающих многоуровневость их структурной организации, взаимопроникновение детерминированности и случайности, определенной автономии элементов данных систем и гармонии целого. Все это привело к качественному изменению в способах научного подхода к изучению таких систем, к изменению логической структуры теорий. В конечном итоге изменяется стиль мышления исследователей-ученых, занимающихся науками о человеке. Он становится инженерно-математическим, с одной стороны, и вероятностным – с другой [1÷5]. Мы назовем его инженерно-вероятностным.

Являясь дальнейшим развитием естественно-научных представлений классической науки, основанных на принципах жесткой детерминации, инженерно-вероятностный стиль научного мышления отличается «квантованием» информации, получаемой в любой области знания и воспроизведением ее в виде целостной системы с помощью средств математической логики; отказом от абсолютизации идей, представлений, методов и форм познания, характеризующихся принципом однозначной детерминации, и более строгим пониманием причинности, включающим в себя принципиально-вероятностные связи; моделированием процессов различной природы методами дискретной математики; реализацией моделей в технических устройствах,

Все сказанное в определенной степени справедливо и для современного лингвистического стиля мышления. Те новые задачи, которые поставила жизнь перед лингвистикой, уже не могут быть решены с точки зрения системно-структурного взгляда на язык и его функции. Установление между элементами современного общества более сложных качественных и количественных связей в жизни общества повышает роль управления, информации и информационных процессов. К таким же системам относится и текст. В нем переплетаются различные аспекты: структурно-коммуникативный и номинативный, социальный и индивидуальный, статический и динамический, детерминированный и вероятностный и т.д. Только учитывая эти факторы можно раскрыть закономерности, которые руководят переходом некоторого «мысленного образования», содержания текста, в цепочки взаимосвязанных предложений, т.е. в текст. Последний может быть изучен или путем анализа, разложения на составные части, или с помощью синтеза из составляющих текст единиц. При втором подходе выявляются те закономерности, которые действуют при объединении слов в предложения, предложений в абзацы и т.д. Знание таких закономерностей поможет не только по-новому подойти и проблемам машинного перевода, автоматического реферирования, аннотирования и поиска текстов, но и создать процедуры синтеза различных стандартных документов. Именно такими исследованиями проверяют «работоспособность» тех правил и исключений, которые действуют на разных уровнях функционирования языка. Подобное моделирование необходимо в лингвистике потому, что лингвист не имеет непосредственно дела с предметом своего исследования – языком. В процессе построения порождающих моделей выполняются те отношения, которые существуют между системой языка, его нормой, узусом и речью [6÷8]. Подход к тексту с точки зрения его порождения позволяет раскрыть некоторые закономерности, остающиеся невыясненными при других методах анализа [9].

Процедура вероятностно-алгоритмического порождения текста основывается на следующих основных идеях:

1. Текст является отражением некоторой жизненной ситуации, воспринимаемой с точки зрения конкретного автора. Однако сам процесс восприятия обусловлен социально, т.е. любой автор членит действительность, а с ней и текст, на социально значимые «куски». Минимальным смысловым «куском» или «квантом содержания» является абзац, показывающий некоторую микроситуацию.

2. Основой или темой текста служит некоторое число опорных слов и словосочетаний, которые отражают главные субъекты, объекты, основное место и время событий, описанных в тексте. Они являются своеобразными смысловыми вехами всего текста и объединяются в таблицу основного статистического содержания текста (ТОСС).

3. Динамика текста понимается нами как «приведение в действие» его опорных слов, т.е. последовательная цепь их отношений между собой и с другими словами в разных частях текста. Наименьшей единицей

динамической составляющей текста является семантико-синтаксическая формула абзаца (СЕСФА), отражающая указанные отношения между составляющими некоторой микроситуации внутри всей ситуации, описанной в тексте. Каждая СЕСФА связана последовательностью семантико-синтаксических формул предложений (СЕСФП), записанных на специальном языке СЕМСИНТ.

4. Конкретизация отношений главных субъектов и объектов в пределах микроситуаций, а также уточнение прочих составляющих микроситуации, описанной в абзац, осуществляется таблицей слов-конкретизаторов (ТАСК), которой снабжается каждая СЕСФА. ТАСК строится на основе ролей, выполняемых словами абзацев в СЕСФП,

5. Текст с заданной ТОСС представляется а виде семантико-синтаксической формулы текста (СЕСФТ), состоящей из последовательности СЕСФА, типичных для данного автора или рассматриваемой предметной области.

6. Подбор СЕСФА для заданного содержания и порядок их следования в пределах СЕСФТ определяется двумя типами факторов: вероятностными, выявленными в процессе статистического анализа следования абзацев с разным, предметно-логическим содержанием а текстах, конкретной предметной области; детерминированными, полученными в результате качественного изучения взаимосвязи ТОСС, ТАСК и СЕСФА.

Главными при этом являются детерминированные правила.

7. Лексическое наполнение каждой СЕСФА проводится также с проверкой двух типов условий: детерминированных, определяемых конкретным содержанием ТОСС, ТАСК и однозначными кодами местоимений, предлогов, союзов и союзных слов, частиц и логико-смысловых скрепов, а также связанных с вероятностным выбором прилагательных, причастий, наречий из списков словосочетаний, типичных для данной области знаний.

8. Грамматическое оформление каждого предложения, входяшего в СЕСФА, осуществляется по строго детерминированным правилам, опирающимся на синтаксические значения семантических падежей, лежащих в основе языка СЕМСИНТ, и грамматические значения предлогов и слов всех других классов. Видо-временной аспект текста задается указанием при словах-конкретизаторах соответствующей грамматический информации.

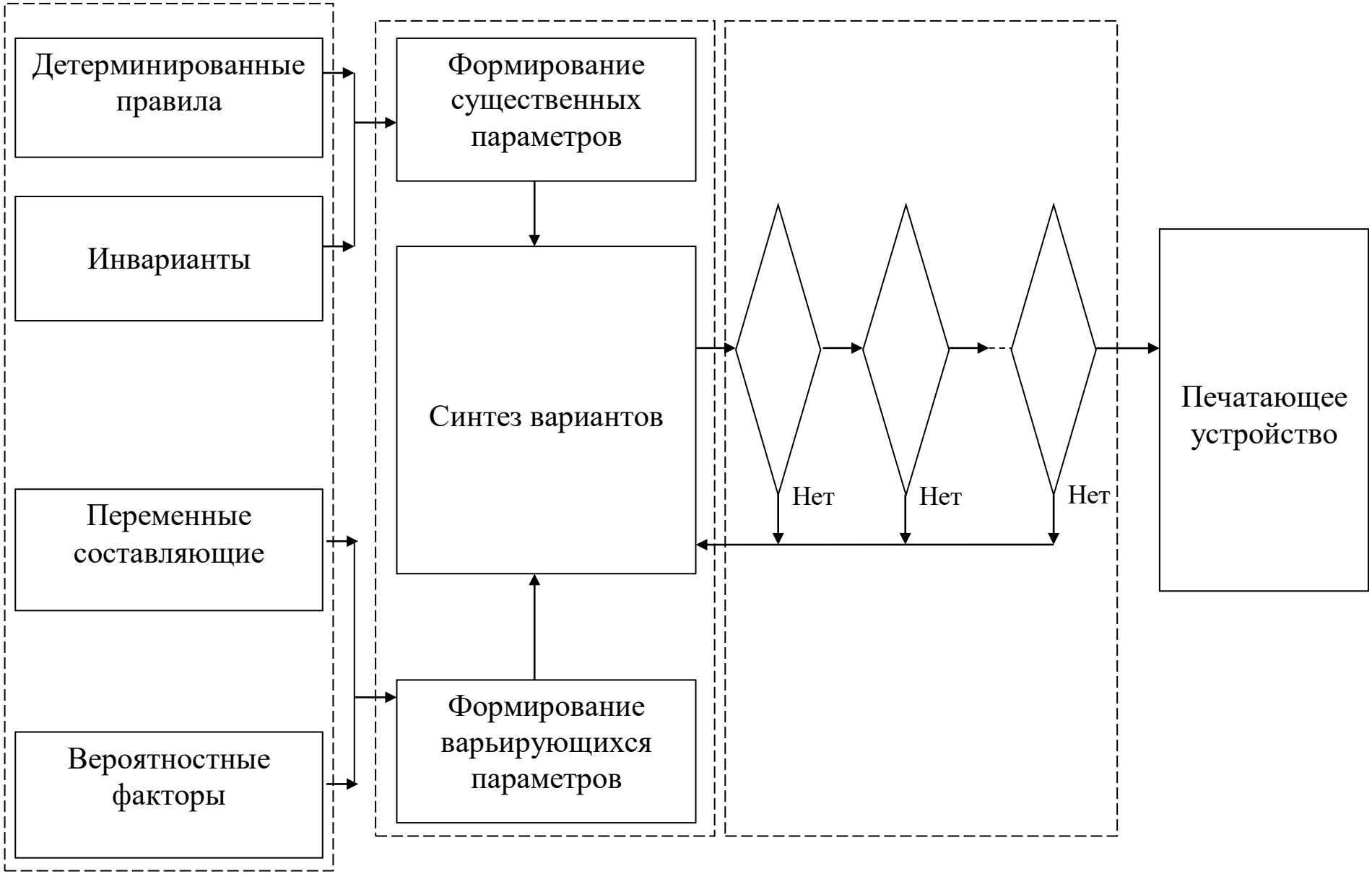
Реализация этих идей показана на схеме (см. рисунок).

Часть I этой схемы связана с выявлением инвариантов текстов определенной предметной области. Части II и Ш представляют процесс порождения текста. В части II выбираются очередные лингвистические единицы (СЕСФА, слова из ТОСС и ТАСК, единицы заполнения из ЛЕ-сочетаний) с учетом как детерминированных правил, так и случайных факторов. Соответствие выбранных единиц определенным требованиям проверяется в части III. В итоге в памяти ЭВМ получается некоторый текст (или его фрагмент), записанный на внутреннем языке машины.

«Расшифровка» его осуществляется при выводе на печатающее устройство или дисплей.

Для реализации этой схемы и построения детальных алгоритмов, выявления наиболее употребительных слов, словосочетаний, СЕСФП, СЕСФА и ТАСК проведен анализ трех видов русских и казахских текстов. Русские тексты охватывают научные, публицистические и поэтические стили, а казахские – художественные, публицистические и научно-популярные.

Программа порождения текста состоит из двух самостоятельных частей: вспомогательной программы, записывающей во внешнюю память ЭВМ всю необходимую лингвистическую и энциклопедическую информацию; основной – реализующей разработанный нами алгоритм порождения текста,


Детерминированные правила
Инварианты
Переменные составляющие
Вероятностные факторы
Формирование существенных параметров
Синтез вариантов
Формирование варьирующихся параметров
Нет
Нет
Нет
Печатающее устройство

Обе программы пишутся на алгоритмическом языке ПЛ/1 для ЭВМ ЕС-1045.

Рассматривая проблему порождения текста в более широком плане, следует отметить ее особую значимость дня теории, и практики создания систем искусственного интеллекта. Опыт их функционирования показывает, что повышение эффективности таких систем возможно только путем использования для человеко-машинного общения естественного языка [10; 11].

Литература

1. Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений. М., 1970. 165 с.
2. Налимов В.В. Вероятностная модель языка. М., 1974. 272 с.
3. Новик И.Б. Вопросы стиля мышления а естествознании. М., 1975. 144 с.
4. Пиотровский Р.Г. Инженерная лингвистика и теория языка. Л. 1979. 112 с.
5. Румянцева Э.А. Инженерно-математический стиль мышления в современной науке. Минск, 1978. 150 с.
6. Алексеев П.М. Квантитативные аспекты речевой деятельности//Языковая норма и статистика. М., 1977. С.43–58.
7. Бектаев К.Б., Белоцерковская Л.И., Пиотровский Р.Г. Норма–ситуация–текст и лингвистические приемы исследования //Языковая норма и статистика. М., 1977. С.5-42.
8. Пиотровский Р.Г., Турыгина Л.А. Антиномия «язык-речь» и статистическая интерпретация нормы языка //Статистика речи и автоматический анализ текста. Л., 1971. С.5-46.
9. Лотман Ю.М. Стихотворения раннего Пастернака и некоторые вопросы структурного изучения текста //Труды по знаковым системам. IV Уч. зап. Тартуск. гос.ун-та. Вып. 236. Тарту, 1969. С.206-238.
10. Вейшенбаум Дж. Возможности вычислительных машин и человеческий разум. От суждений к вычислениям. М., 1982. 368 с.
11. ЭВМ пятого поколения. Концепции, проблемы, перспективы. М., 1984. 110 с.

М.ӘУЕЗОВ ТІЛІНІҢ СӨЗДІКТЕРІН ЖАСАУ ТӘЖІРИБЕСІ

Әл-Фараби атындағы Қазақ мемлекеттік
ұлттық университеті. ҚАЗМУ «Хабаршысы».
Филология сериясы Алматы-1997, №8. 18-24-бб.
(С.Мырзабековпен бірге)

Мұхтар Омарханұлы Әуезов – аса көрнекті қазақ жазушысы, ұлы суреткер, әдебиетші-ғалым, зор қоғам қайраткері, КСРО кезіндегі Лениндік және Мемлекеттік сыйлықтарының лауреаты.

Жазушының шоқтығы биік үздік шығармасы «Абай жолы» роман-эпопеясының тілі лингвистикалық статистика тұрғысынан біршама зерттеліп, одан жиілік сөздіктер (1979 ж.) жасалып, романда 16983 сөз (жалқы есімді қосып есептегенде) 61424 формада (сөзформа) 465691 рет қайталанып (сөзқолданыс) қолданылғаны анықталды.

1979-1985 жылдары Қазақстан Республикасы Ұлттық Ғылым академиясының М.О.Әуезов атындағы Әдебиет және өнер институтының жазушы мұрасын 20 том етіп жарыққа шығарды. Мұнда ұлы жазушының негізгі көркем туындылары, аудармалары, ғылыми зерттеу, сын және публицистикалық еңбектері, әр алуан мәжілістерде, әдеби айтыс жиындарда сөйлеген сөздері мен баяндамалары жарияланды.

М.Әуезовтің 20 томдық шығармалар жинағы бойынша әр түрлі сөздіктер алу жұмысын ҚР ҰҒА-ның А.Байтұрсынұлы атындағы Тіл білімі институтының Статистика тобы 1986 жылы қолға алған болатын. Сол жұмыстың нәтижесінде «М.О.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері» (22 баспа табақ) 1995 жылы жарық көрді. Сөздік қазіргі заман компьютерлерінің көмегімен алынды.

Белгілі бір жазушының немесе жеке бір шығарманың тіл байлығы, сөз қолдану шеберлігі кім-кімді де бей-жай қалдырмаса керек. А.С.Пушкин өз шығармаларында 21197, В.Шекспир 15000, Абай 6000-нан аса сөз қолданғаны белгілі.

М.Әуезовтің 20 томдық шығармалар жинағында қанша сөз қолданғаны, яғни жазушының тіл байлығы жөнінде ендігі жерде нақты айтуға болады. Сол үшін де қыруар жұмыс атқарылып, сөздік жасалды.

Сөздік жасау жұмысы кезең-кезеңмен жүргізіледі. Соның алғашқысы сөздерді таптастырудың, сөз табына ажыратудан басталады.

Қазақ тілі омоним сөздеріне бай. 10 томдық түсіндірме сөздікте (қараңыз: Қазақ тілінің түсімдірме сөздігі. Алматы, "Ғылым", 1976-1986 ж.) 7 мыңдай омоним бары анықталды. Мәселен, *ай* – зат есім, *ай* – шылау, *ай* – одағай; *ақ* – сын есім, *ақ* – зат есім, *ақ* – етістік, *ақ* – шылау; *ау* – зат есім, *ау* – етістік, *ау* – одағай, *ау* – шылау т.б. Тексте бұлардан басқа *қалам* – зат есім, *қалам (қала-м)* – зат есім, *қаламын (кал-а-мын)* – етістік, *қала-сын (қала-сын)* зат есім, *қаласын (қала-сын)* – етістік сияқты толып жатқан сәйкестіктер омоним формалар ұшырайды. Оларды алдын ала саралап, таңбалап бермесе, компьютерге арнайы бағдарлама жасау қиындыққа түседі. Сондықтан тексте кездесетін әрбір сөзқолданыстың (сөздің) тұсына олардың қай сөз табына қатысын айқындап, таңбалап шығуға тура келеді.

Сонда 20 томдықта 2,1 млн. сөзқолданыс (сөз) кездессе, сонша рет әрқайсысының тұсына сөз табы көрсетіледі деген сөз. Сөздерді топтастырудың да өзіндік қиындығы бар екені тілші қауымына белгілі. Тілдегі сөздердің сөз табына қатысын айқындау, әр сөз табының ауқым-аумағын, қолданылу ерекшелігін білудің теориялық та, практикалық та

маңызы зор. Бұл салада қазақ тіл білімінде біршама статисткалық зерттеу, мағлұматтар, тәжірибе жинақталды. Солардың алғашқыларының қатарында М.Әуезовтің «Абай жолы» эпопеясында қолданылған сөз, сөзформа сөзқолданыс сөз табына ажыратылып, статистикалық тұрғыдан сипатталды («М.Әуезовтің «Абай жолы» романының жиілік сөздігі», 1979).

Одан бергі жерде тіліміздегі сөздерді түгел қамтуды мақсат еткен 10 томдық «Қазақ тілінің түсіндірме сөздігі» дүниеге келді. Мұнда тізбе сөздердің әрқайсысының сөз табына қатысы көрсетілген. Мұның өзі оларға статистикалық талдау жасауға мүмкіндік берді (Қазақ тексінің статастикасы II шығарылуы, А., 1990, 126 б.).

Сөздерді топтастырудың принциптері айқын болғанымен, тілдгі әрбір сөздің грамматикалық табиғатын жазбай тану іс жүзінде қиынның қиыны. Сол қиындықтан шығу үшін дүдәмал сөздер тобын пайдаланған болатынбыз. Бұған *модаль, әртарап, бейтарап* делініп жүрген сөздер жатқызылған-ды. Сондай-ақ сөз табына қатысын тап басып тануға қиын сөздер бұрынғы тәжірибемізде осында топтастырылған еді.

20 томдықтан сөздік алу барысында бұл тәжірибеден бас тартып, тіліміздегі кез келген сөз оқулық зерттеулерде баяндалатын тоғыз сөз табының біріне жатқызылды. Сөз таптары қысқартылып шартты таңбамен белгіленді: зат есім – *зт.,* етістік – *ет.,* сын есім – *сн.,* есімдік – *ес.,* үстеу – *үс.,* одағай – *од.,* сан есім – *са.,* еліктеуіш сөздер – *ел.*

Жазушы тілінің лексикалық қабаттарын неғұрлым саралап, нақты білу үшін бұлардан басқа да шартты таңбаларды пайдалануға тура келді. Олар: *бн.* – кірме сөздер, *А* – адам аттары, *Г*– жер-су аттары, *Р* – ру аттары, *Ж* – жан-жануар аттары, *Б* – баспа аттары, *Н* – мекеме аттары, *С* – қысқарған сөздер, *Д* – атақ-даңқ, лауазым атаулары. Бұлардың ішінен сөздікте тек *бн.*– кірме сөздер ғана орын алды.

Осы кезеңге жататын бір өте маңызды жұмыс – 20 томдық шығармалар жинағының тексін компьютер жадына енгізу. Бұл жұмыс та қатаң қадағалаудың және бірнеше рет тексеру жүргізудің арқасында іске асты.

Екінші кезеңде әрбір томнан алынған алфавитті-жиілік сөздіктердің негізінде *үлестірімді алфавитті-жиілік сөздік* алынды. Бұл аталған сөздікте 20 томдағы барлық сөзтұлғалардың алфавит тәртібімен орналасқан тізімі берілді және әр томдағы кездесу жиіліктері бөлек-бөлек көрсетіліп, соңынан олардың 20 томдағы қосынды жиілігі шығарылды. Мұндай сөздіктің теориялық та, практикалық та маңызы зор, сондықтан да болашақта жеке сөздік ретінде шықса, зерттеу жұмысында ғана емес, оқыту барысында да пайдасы тиер еді деген үміттеміз.

Үшінші кезеңде үлестірімді алфавитті-жиілік сөздіктің негізінде сөзтұлғалар тізбе сөз түріне келтіріліп, *алфавитті-жиілік сөздік, жиілік сөздік* және *кері алфавитті-жиілік сөздіктер* алынды.

Алфавитті-жиілік сөздікте сөздер алғашқы әрпінен соңғы әрпіне қарай қатаң алфавит тәртібімен орналасса, *кері алфавитті-жиілік сөздікте* – керісінше сөздер соңғы әріптерінен бастап қатаң алфавит тәртібімен орналасады.

Жиілік сөздікте сөздер олардың текстегі кездесу жиілігінің біртіндеп кему ретімен беріледі, яғни сөздікте бірінші орында тұратын сөз – жиілігі жағынан ең көп кездесетін сөз, ал одан кейінгілері жииліктерінің біртіндеп кему ретімен орналасады да жиілігі бірдей сөздер қатаң алфавит тәртібімен тізіледі.

Аталған сөздіктердің маңызына және олардың кейбір сандық (статистикалық) мәліметтеріне қысқаша тоқталайық.

Алфавитті-жиілік сөздікте өзінің мақсаты мен міндетіне қарай әрбір сөздің лексикалық мағыналары жөнінде толық түсінік берілмейді, тек грамматикалық мағыналары (сөз табы) ғана анықталып, текстегі кездесу жиілігі көрсетіледі.

Алфавитті-жиілік сөздік зерттеуге қажет болған сөздің қолданылу (оның туынды формаларын да) жиілігін табуға, зерттеуге, М.Әуезовтің толық шығармалар жинағының лексикасының жалпы құрылымын анықтауға мүмкіндік береді. Мұндай сөздік нормативті сөздік емес. Өйткені 20 томдық жинақта сөз бен оның туынды формалары қалай жазылса, сөздікке солай енгізілді. Сондықтан кейбір сөздердің жазылуында әдеби нормадан ауытқу кездесуі мүмкін. Солай бола тұрса да, тіл мәдениетін зерттеушілер бұл сөздіктен бағалы материалдар таба да, авторлың тіл ерекшеліктерін байқай, анықтай да алады. Морфологиялық, фонетикалық дублет сөздердің автор қандай нұсқаларын жиі не сирек қолданғанын сөздіктен толық анықтап, тілдегі жарыспалы сөздердің (дублеттердің) белгілі бір нұсқасын әдеби норма етіп алуға алфавитті-жиілік сөздік үлкен көмегін тигізеді. Сол сияқты оқушы-зерттеуші осы сөздіктен М.Әуезовтің қай сөзді қандай тұлғада неше рет қолданғанын да біле алады.

Алфавитті-жиілік сөздікте жалпы тіл біліміне қатысты басқа да бірнеше теориялық мәселелердің бетін ашуға, М.Әуезов шығармаларының тіл ерекшеліктерін анықтап, түсіндірме сөздігін жасауға септігін тигізеді.

Алфавитті-жиілік сөздікте тізбеге енген сөздердің жалпы саны – 29483 (жалқы есімді қоспағанда). Енді осы сөздердің бастапқы әріптеріне қарай сөздікте орналасуының сандық мәліметтерін М.Әуезовтің 20 томдық шығармалар жинағынан алынған алфавитті-жиілік сөздік бойынша қарастырайық. Ең алғаш байқайтынымыз, сөздердің бастапқы әрпіне қарай сөздің саны да өзгеріп отырады. Бұл заңдылық барлық тілдерге тән құбылыс десе де болады.

Мысалы, М.Әуезовтің шығармалар жинағынан алынған алфавитті-жиілік сөздікте жеті әріптің әрқайсысынан басталатын сөздер саны екі мыңнан асады. Олар «К»-2398, «Т»-2817, «С»-2431, «Б»-2370, «Ж»-2327, «А»-2306, «К»-2219 тізбедегі сөздердің саны 29483 десек, осы 7 әріптен басталатын сөздер жазушы тілінің 59%-ін құрайды екен. Тек қана «Қ» және оның жіңішке варианты «К» әрпінен басталатын сөздер тізбенің 18%-ін, «Т»-9,5%-ін, ал қалған 5 әріптің әрбіреуі тізбенің (реестрдегі) 8%-ін қамтиды.

Енді осы 7 әріптен басталатын сөздердің текстегі қайталануын ескерсек, сөзқолданыстың қанша процентін қамтитынын есептеп шығаруға

болады. Жалқы есімсіз текстің «ұзындығы» (көлемі) 1706195 сөзқолданыстан тұрады. Сонда «Қ» әрпінен басталатын сөздердің қосынды жиілігі – 148337, ол текстің 8,7% процеті. Сол сияқты, «Т»дан басталатын сөздер 125989 рет қайталанып, текстің 7,4%-ін қамтиды, «С»- 23166 (7,2%), «Б»- 237698 (14%), «Ж» 165448 (10%), «А»152663 (9%), «К»- 135894 (8%). Сонымен, аталған 7 әріпттен басталатын сөздер қайталана келе М.Әуезовтің 20 томдық шығармалар жинағы тексінің 64,3%-ін қамтиды екен.

Бұл қысқаша мәліметтердің өзі М.Әуезовтің тіл ерекшетігін білдіретін деректер, ал басқа жазушының шығармаларында мұндай деректер өзгеше болуы да мүмкін.

Енді жиілік сөздіктің сандық сипаттамасына тоқталайық.

Сөздердің орналасуы алфавит тәртібімен емес, қолданылу жиіліктерінің бірте-бірте кему жолымен берілген сөздер (лексикалық единицалар) тізімін жиілік сөздік деп атаймыз.

Тізбедегі 29483 сөз (жалқы есімді қоспағанда) жиіліктерінің кему ретімен орналасқан.

Жиілік сөздікте бірінші орында «бол» етістігі тұр. Ол М.Әуезовтің шығармалар жинағының тексінде 38944 рет қайталанып кездеседі екен, бұл барлық текстің 2,28%-ін құрайды. Екінші орында «де» етістігі тексте 31901 рет кездеседі (1,87%). Осы екі етістік барлық текстің 4,15%-ін қамтитынын байқауға болады. Мұның өзі көмекші мағынада көп қолданылатын «бол», «де» етістіктерінің жұрнаққа айналу қарсаңында тұрғанын байқатады. Сөздіктегі осы сияқты ең жиі кездесетін әр түрлі сөздердің барлық сөзқолданыстың қанша процентін қамтитынын жиілік сөздіктен көруге болады. Тек қана тізбедегі алғашқы 10 сөз (*бол*/ет., *де/*ет., *бұл/*ес., *е*/ет., *өз/*ес., *да/*шл., *ал/*ет., *кел*/ет., де/шл., ол/ес.) барлық текстің 13%-ін камтиды. Ал осы сияқты сөздіктің бас жағында тұрған 31 сөз текстің 1/4-ін яғни 25%-ін, 174 сөз -1/2-ін, яғни 50%-ін, 1292 сөз - 80%-ін қамтитынына көз жеткізуге болады. Бұл сөздер қандай сөздер, грамматикалық лексикалық мағыналары қандай, неге тексте осы сөздер жиі кездеседі, М.Әуезовтің тіл ерекшелігіне жата ма, әлде жалпы қазақ тіліне тән қасиет пе деген сұрақтарға жауап беру үшін арнайы зерттеулер қажет.

Жиілік сөздіктегі сөздер жиілігіне сәйкес топ-топқа бөлініп, олардың реестрдегі сөздердің қанша процентін және қайталану қабілетін ескеріп текстің (сөзқолданыстың жалқы есімсіз) қанша процентін қамтитынын байқауға болады. Мысалы, жиілігі 10000-нан асатын 17 сөз тізбедегі барлық сөздердің 0,0006%-ін қамтығанымен, қайталана келе бұған текстегі сөзқолданыстың 18,26%-ін құрайды. Сол сияқты, жиілігі 501-ден 10000-ға дейінгі сөздердің саны 30-ға тең тізбедегі 29483 сөздің бұл 0,001% проценті, ал барлық, текстің (қайталануын ескергенде) небәрі 0,122%-і екен. Жиілігіне сәйкес басқа топтардың қамту проценттерін осылайша сөздік бойынша анықтау қиын емес.

Жиілік сөздіктің негізінде барлық тілдерге (қазақ тіліне де) тән заңдылықты байқауға болады. Ол – сөздердің жиілігі кеміген сайын реестрдегі сөз санының ұлғаюы, бірақ текстегі сөз қолданысты қамту

проценті барынша аз болуы. Бұл заңдылықты кестедегі жиілігі 10-нан 1-ге дейін кеміп отыратын сөздердің санынан анық байқаймыз. Мысалы, жиілігі 10-ға тең сөздердің саны сөздікте 378 (0,013%) болса, жиілігі 9-ға тең сөздердікі – 465 (0,016%), Осылай жиілігі кеміген сайын олардың қатары арта түсетінін аңғарамыз. Мәселен, жиілігі 2-ге тең сөз саны 3712 (0,126%), ал жиілігі 1-ге тең болғанда – 10 988 (0,373%) дейін өсетінін заңды құбылыс деп есептеуіміз қажет. Мұндай төменгі жиілікте қолданған сөздер жазушы тілінің мәйегі деуге болады.

Көбінде жазушының сөз байлығы осындай сирек кездесетін сөздердің санымен сипатталады. Шынында да Мұқаңның тың сөз қолдану шеберлігі оқырмандарға бегігілі, сонымен қатар бір қолданған сөзін екінші рет кайталамай, баламасын іздеп, омоним, дублет сөздерді көп қолданатыны да сөз байлығын көрсетеді. Әрине, алдымызда басқа да белгілі сөз шеберлерігін сөздіктері жатса, ондағы осы құбылыстарды айқындап салыстыра зерттесе, әр автордың өзіне ғана тән тіл ерекшеліктерін нақты алуға болар еді. Мұндай зерттеулер болашақта дүниеге көптеп келеді деген сенімдеміз. Әзірге қолда бар там-тұм текстер бойынша алынған деректерді келтірейік.

Абай тілінің жиілік сөздігі бойынша Абай шығармалары тіліндегі ең жиі қолданған 75 сөз барлық сөзқолданыстың 40,8%-ін, 150 сөз – 50,8%-ін, 1000 сөз – 80%-ін құрайды. Ал қазақ тілінің газет текстерінен жасалған жиілік сөздіктің негізінде (қараңыз. А.Ахабаев Қазіргі қазақ газеттері тілінің алфавиттік-жиілік сөздігі, «Қазақ тілінің статистикасы», Алматы, «Ғылым», 1973) ең жиі қолданылған 57 сөз жалпы сөзқолданыстың 25%-ін, 301 сөз – 50%-ін, 1168 сөз – 75%, 1602 сөз – 80%-ін қамтиды.

Енді сирек кездесетін сөздерге келсек, Абай тіліндегі бір реттен жұмсалған 2975 сөз ақын қолданған барлық сөздің 49,4%-ін, қайталануын ескергенде жалпы сөзқолданыстың 6,35%-ін қамтитыны анықталды. Зерттеуге алынған газет тексінде бір реттен қолданылған 5720 сөз жалпы сөздердің (12424) 46,04%-ін, ал барлық сөзқолданыстың 3,81%-ін құрайды.

Бұл мәліметтер публицистика тілі мен көркем шығармалар тілінің ерекшеліктерін аңғартып қана коймай, Абайдың да, Мұқаңның да аса зор көркем сөз шеберлері екенін, олардың тіл байлығын танытатын, дәлелдейтін деректер.

Жиілік сөздердің маңызына тоқталсақ олар ең алдымен тілдің лексикалық құрамын терең зерттеуде, лексикалық жалпы функционалдық стильге ортақ бөлігін ажыратуға мүмкіндік туғызады. Сондай-ақ жалпы халықтық лексиканың кірме сөздер мен терминдердің, неологизмдердің және басқа лексикалық топтардың тілдік жалпы лексика жүйесінен алатын орны жөнінде көптеген мәлімет алуға жәрдемдеседі.

Сөз болып отырған *жиілік сөздік* М.Әуезов тілінің лексикалық байлығы мен стильдік ерекшелігін, сөз қолдану шеберлігін анықтауға және оларды туыстас түркі тілдерімен салыстыру, яғни типологиялық зерттеулер жүргізу үшін де аса қажетті материал береді деуге болады.

Жиілік сөздік көпгеген лексикографиялық мәселелерді шешуге объективті негіз бола алады. Мысалы, екі немесе одан да көп тілден аударма

және түсіндірме сөздік жасаушылар бұл Сөздіктен құнды-құнды материал таба алады.

Кітаптағы үшінші Сөздік М.Әуезовтің 20 томдық шығармалар жинағының *кері-алфавитті жиілік сөздік* (кейде *кері-алфавитті жиілік сөздік).* Бұл сөздікте сөздердің орналасу тәртібі соңғы әріп-дыбысынан бастап қатаң алфавит тәртібімен берілген. Сол себептен сөз соңындағы сөз таптарын білдіретін шартты белгідердің әсерінен әр сөз табына жататын топтар бөлек-бөлек беріледі. Сондықтан сөздікте әр сөздің соңына шартты белгілер көрсетілмей «Зат есім сөздер» деп атап бірінші топ, «Етісік сөздер» – деп екінші топ сөздер, т.с. рет-ретімен берілді.

Лексикография тарихында *кері-алфавитті сөздіктің* екі түрі бар. Біріншісі – бұрын қарастырылған түсіндірме не екі тілді сөздіктің негізінде жасалған *кері-алфавитті сөздік.* Сөздіктің бұл түрінде сөз және туынды сөздер (сөз тудырушы формалар) бар да, ал сөзтұлғалар (сөзтүрлендіруші морфемалар жоқ және сөздердің қолданылу жиіліктері де көрсетілмейді.

Кері-алфавитті сөздіктің екінші түрі – *кері-алфавитті жиілік сөздік* деп аталады. Сөздіктің бұл түрінде де сөзтұлғаның зерттеуге алынған текстерде қанша рет қолданғаны көрсетіліп отырады.

М.Әуезовтің 20 томдық шығармалар жинағының тексінен алынған кері-алфавитті жиілік сөздіктің реестр сөздерінің негізіне «сөз» және сөз тудырушы формалар (туынды сөз) алынды. Сондықтан бұл сөздікте әр сөз табына қатысты сөз жасаушы жұрнақтар бір жерге жинақталып беріледі.

Кері-алфавитті жиілік сөздіктер морфологиялық категорияларды зерттеу үшін өте құнды материалдарды түгел және тез табуға мүмкіндік туғызады. Бұл сөздерден зерттеуге алынған текстердегі сөз тудырушы және сөз түрлендіруші морфемалардың құнарлы не құнарсыз екендігін анықтау үшін, терминдер мен кірме сөздерге жалғанған қосымшалардың құрылымын талдау үшін бағалы-бағалы тілдік деректерді алу әрі оңай, әрі ыңғайлы.

Кері-алфавитті жиілік сөздіктің әрбір сөз табының қолданылу жиіліктерін және сол сөз табына қатысты әрбір жұрнақтың (жалғаудың, олардың нұсқаларының) қолданылу жиіліктерін анықтауға мүмкіндігі мол.

Енді 2.1 млн. сөзқолданыстан тұратын М.Әуезовтің 20 томдық шығармалар жинағындағы сөз таптарының сөздік пен текст арасындағы салыстырмалы статистикалық деректеріне қысқаша тоқталайық.

1-кестедегі сандық көрсеткіштен тексте ең көп қайталанып қолданылатын сөз таптары алдымен етістік; екінші орында – зат есім, ал ең аз кездесетіндер одағай мен еліктеу сөздер екенін аңғаруға болады.

Осы жерде А.Толстойдың «көркем шығарманың жаны – етістіктер» дегені еске түседі.

Текстегі сөзқолданыстың 18%-ін құрайтын жалқы есімдерді қоспағанда М.Әуезовтің 20 томдық шығармалар жинағында 29483 сөз қолданылғаны аннықталды. Оның 5743-і немесе 19,48%-і кірме сөздер (бн). Мұндағы кірме сөздер негізінен орыс тіліндегі орфограммасын сақтап енген сөздер және солардың өрбіген туынды түбірлер. Бұлардың қатарында қазіргі орфографиялық нормаға сәйкес келмейтін араб-парсы сөздері де бар.

Мысалы, фарсы, фиғыл, фарасат, фақыр, фәле т.б. Орыс тілі де әртүрлі фонетикалық нұсқада ұшырайды. Мысалы, прокурор – пыркорол, протокол – пүртөкөл, генерал – жанарал, приговор – пригоуор – пірғауар, т.б. Кейіпкерлердің тіліндегі айтылу ыңғайына орай жазылған әртүрлі фонетикалық нұсқадағы сөздерге де тізбеде орын берілді. Мысалы, ажуа (55 рет қайталанып қолданылған), әжуа – 22, әжуала – 7, ажым – 65, әжім – 15, ажымды – 11, әжімді – 4 және т.б.

1-кесте

Қ/С	Сөз табы	Сөз саны	Тізбедегі үлесі (%)	Сөз қолданыс саны	Текстегі үлесі (%)
1.	зат есім (зт)	8843	37,25	520521	30,50
2.	етістік (ет)	6853	28,87	534836	31,35
3.	сын есім (сн)	5885	24,79	189624	11,00
4.	есімдік (ес)	209	0,88	173495	10,7
5.	сан есім (са)	94	0,40	28009	1,64
6.	үстеу (үс)	1023	4,31	74472	4,36
7.	шылау (шл)	162	0,68	129536	7,60
8.	одағай (од)	327	1,38	8408	0,50
9.	еліктеуіш (ел)	344	1,45	2029	0,12
кірме (бн) сөздер		5743	19,40	45265	2,65
Қосындысы		29483		1.706195	

Бұл кестеден М.Әуезов сөз байлығының үштен екісі зат есім, етістік сөздер екенін байқаймыз. Оның есесіне есімдік пен шылау сөздер бірігіп жазушы тілінің жарты процентін құрайды, алайда олардың тексте қайталану жиілігі өте жоғары, барлық сөзқолданыстың 18%-іне жуықтайды.

Зат есім тудыратын ең өнімді *-шы, -ші* жұрнағы 343 сөздің құрамында ұшырасса, етістік тудыратын *-ла/-ле, -да /-де, -та/-те* 1409 сөз жасап тұр.

Жазушы тілінде кездесетін сын есім сөздердің жартысынан көбі (3054 сөз) *-дай/-дей, -тай/-тей, -лы/-лі, -ды/-ді, -ты/-ті, -сыз/-сіз* жұрнақтары арқылы жасалған.

М.Әуезовтің шығармалар жинағында жалқы есімдер 386729 рет қайталанып қолданылған. Бұл сөздер сөздіктің реестр қатарында жоқ, бірақ әр том бойынша және түгел 20 том ішіндегі жалқы есім жайлы мәліметтер бар. Жалқы есімдер іштен сегіз түрге бөлініп тексте шартты белгілермен таңбаланады.

386729 рет қайталанған 20 томдық шығармадағы жалқы есімдер жайлы мәліметтер төмендегідей: адам аттары – 173300 рет, жер-су аттары – 115396 рет, ру-тайпа аттары – 89705 рет, жан-жануар аттары – 3813 рет, баспа аттары – 3345 рет, мекеме аттары – 592 рет, қысқарған сөздер – 526 рет және атақ-даңқ аттары –12 рет.

Бұл мәліметтерден бірден көзге түсетіні – текстің өн-бойында адам аттары мен есімдік сөздердің қайталану санының (173300; 173495) жуықтас

болуы және атақ-даңқ, лауазым атауларының сирек қолданылып, тек мақалалар мен зерттеулерде ғана аздап кездеседі.

Сонымен, М.Әуезовтің 20 томдық шығармалар жинағында қайталанатын әртүрлі сөз таптары мен атаулардың сан мөлшерінің өзіндік қыры-сыры, тілдік таным-түсінігі бар.

Қорыта келе айтатынымыз, ұсынылып отырған үш түрлі жиілік сөздіктер М.Әуезовтің тіл байлығы мен стильдік ерекшеліктерін тереңірек зерттеу үшін аса құнды материал болатындығына зор үміттеміз.

Жеке қалам шеберлерінің сөздігімен қатар, қазақ тілінің сөз байлығын, яғни қазіргі қазақ тілінің жалпы лексикасын түгел қамтитын жиілік сөздік құрастыру да күн тәртібінде тұрған мәселе – заман кажеттілігі. МӘуезовтің 20 томдық шығармалар жинағының жиілік сөздіктері жазушының ғана тіл байлығын сипаттап қоймай, қазақ тілінің сөз байлығының көрсеткіші болатын алда құрастырылатын толық жиілік сөздікке сүбелі үлес қосатынына күмән жоқ.

ПРОФЕССОР Қ.ЖҰБАНОВ ЖӘНЕ ҚАЗАҚ ТІЛІН ЗЕРТТЕУДІҢ КЕЙБІР ҚОЛДАНБАЛЫ ТӘСІЛДЕРІ

ҚР Білім, мәдениет және денсаулық сақтау Министрлігі. Қ.Жұбанов атындағы Ақтөбе университеті. Қ.Жұбановтың 100 жылдығына арналған «Қазіргі заманғы гуманитарлық ғылымдарының өзекті мәселелері» атты республикалық «Жұбанов тағлымы» конференцияларының материалдары (1989, 1994, 1996 жж.). Ақтөбе – 1998. 51-56-бб.

Ақтөбе жұртшылығының ардақты жерлесі, қазақ филологиясының тұңғыш профессоры, іргелі, терең ойлы ізденістерімен зиялы жұртшылықты күні бүгінге шейін өзіне баурап алып келе жатқан ғұлама-ғалым Құдайберген Қуанұлы Жұбанов лингвистиканың аты шулы Ленинград мектебін тамамдап, ҚазПИ-ге қайтып оралғаннан кейін ең алдымен ұлт мектептері мен ондағы ана тілі мұғалімдеріне қандай жолдармен ғылыми методикалық көмек көрсетуге болатындығын ойластырған. Профессор, әрі республика Халық Ағарту комиссариатының коллегия мүшесі ретінде «Қазақстанның екпінді мұғалімдерінің бірінші съезінде» [1] сөйлеген сөзінде ол жоғарғы жақтан басталатын «қағазбастылықты» оқыту ісіне есебі тиетін «қағаз басшылыққа» айналдыру қажеттігін айтады. «Руководство можно разбить на живую связь и на руководство бумажного характера, - дей келіп, ол былай деген. - Но если мы в отдельных случаях несколько пренебрегаем бумажным руководством, то от него мы не можем отказаться в деле народного просвещения. Здесь бумажное руководство, наряду с живым, является основным видом руководства, потому то от него мы не можем дать исчерпывающие

методические и организационные указания учителям, по которым они могут построить и перестроить свою работу», - деп көрсетеді [1, 17-22]. Бірақ бұндай нұсқаулар жалпы көпшілікке мазмұны таныс құлаш-құлаш методикалық хат түрінде болмай, мұғалімдердің күнделікті жұмысына көмегі тиетін нақты үлгі-ұсыныс ретінде жазылуы керектігін арнайы ескертеді. Ғалымның өз сөзін аудармай келтірсек, былай деген: «Вместо того, чтобы писать о каких-то общих вопросах, надо давать учителю конкретные указания, как построитъ урок, как лучше преподнести учащимся необходимые знания, как бороться за грамотность, за культуру речи, как оживить урок, как провести экскурсию и т.п. Вот так в дальнейшем надо строить форму бумажного руководства. И на этот путь уже встал Наркомпрос» [1].

Профессор Қ.Жұбанов өзінің бұл айтқандарын жеке басының іскерлік қабілетімен дәлелдеді. Мәселен, ол ҚазПИ-ге келген алғашқы жылынан бастап «Қазақ тілінің ғылыми курсы» бойынша оқыған лекцияларын мұғалімдерге көмекші материал есебінде педагогикалық баспасөз бетінде жариялай бастады. Ғалымның осы мақсатта жарық көрген алғашқы мақаласының бірі деп біз жоғарыда аталған курстың 1933 жылы «Политехннкалы мектеп» журналында басылған бірінші бөлігін айтамыз. Оның «Беташары» былай басталады: «Осы номерден бастап отырған «Қазақ тілінің ғылыми курсы» өткен 1932-33 оқу жылында Қазақ пединститутында (ҚазПИ) оқылған лекциялар еді» [2, 118-134]. Автор бұнда қазақ тілінің фонетикасына, оның ішінде «дыбыс жүйесіне» қатысты өзінің аса маңызды ой-толғауларын ортаға салады. Қазақ тілін алғаш зерттегендер – патша үкіметінің отаршыл саясатын іске асыру мақсатымен отар елдердің әдет-ғұрыпын, тіл-әдебиетін тексерген миссионерлер, одан қалды күншығысшыл оқымыстылар болғандығын айта келіп, өзіне дейінгі және өз тұсындағы тілтану ілімінің тарихи жай-жапсарына байыпты шолу жасайды.

Келесі 1934 жылы ғалым республикалық «Ауыл мұғалімі» журналының (№4) бетінде әрі ғылымдық, әрі практикалық жағынан өте қызықты тағы бір мақала жариялайды. «Буын жігін қалай табуға болады?" [3, 330-340] деп сұрау мәнді тақырып қойылған бұл еңбектің өз кезеңі үшін емес, бүгінгі, тіпті ертеңгі күн үшін маңызы зор. Ғалым-ұстаз өзінің сөзін «Буын жүйесінің маңызын түсіндіруден» бастайды. Ол қазақ тілінің табиғатын дұрыс тану үшін буын жүйесін жете білу керектігін ескертіп: «...буынды жақсы білмесең, емлені де білмейсің; сөз тудыру, сөз түрлендіру сияқты морфология негізін де ұға алмайсың; буынды білмейінше, сөз екпіні (ударение), сөз сазы (интонация) секілді сөйлем жүйесіне қатысты тіл құбылыстарының сырына да қана алмайсың; буынды аңғармасаң, «ажарлап сөйлеу», «әсерлеп оқу» сабақтарын да дұрыс жүргізе алмайсың; буын жайына қанық болмасаң, өлең құрылысының негізімен дұрыс таныс бола алмайсың...» [3] - деп, мұғалімнің зейінін негізгі мәселеге барынша аударып алады да, буын жүйесінің құрылымын, оның алуан тілдердегі атқаратық қызметтерін атап көрсетеді. Әсіресе қазақ тілінде буынның маңызы өзгеше, артықша, өйткені қазақ тілінің дыбыстары «буын ішіндегі дыбыстар болып

бірлесіп қана тұрғылықты түрге енетіндігін» айтады. Буын ішіндегі дыбыстар бірі мен бірі байланысты, бірі бірінің қасы мен қабағына қарап тұрғандай үйірленіп, ымыраласып отырады дей келіп: - Жұртқа мәлім, «үндестік заңы» дейтініміз – дыбыстардың осы ымыраласуы болатын.., - деген ой түйінін ортаға салады.

Осы зандылықтардың сырын тану үшін қазақ сөздерінің статистикалық сипаттамасын жасау үстінде біз мынандай жайттарды анықтадық. Атап айтқанда, автоматты морфологиялық талдаудың алгоритмдерін құрастыру барысында біз туынды сөз тұлғаларының негізгі сөз тұлғаларымен сәйкес келетін тұстарын ескеруге тырыстық. Бұның өзі тілді электронды есептегіш машинамен тексеру ісінде үлкен қиындықтар келтіріп жүрген құбылыс болатын. Мәселен, «бал» (араның балы) сөзі мен «балға» (ұсталық сайман) сөздерін «нанды балға батыр» және «үлкен балға» деген тәрізді қолданыстардан негізгі, туынды түбірлерді автоматты жолмен (санасыз) айырып тану мүмкін болмағандықтан, «ережеге бағынбайтын ерекше сөз тұлғалары» ретіндс өз алдына дербес кестелеп келдік. Сондай-ақ, біздің алгоритімімізде автоматты «іздеу салу» процесін шапшаңдату мақсатымен түбір-негіздің ұзындығы қанша буыннан тұратындығын санау шарты да ескерілген-ді. Бұның өзі қазақ тіліндегі сөз тұлғалардың негізгі түбірін табуды едәуір шапшаңдатуға мүмкіндік беретін. Ал, біз өз алгоритмімізде тек зерттелетін текстегі сөз тұлғалары мен олардың негізгі түбірлерін анықтап қана қоймай, сонымен бірге объектідегі сөз тұлғаларының құрамында сөзді тудырушы және сөзді түрлендіруші қанша қосымшалар тіркесі барын есептейтін «Сөздік» жасауды да мақсат еткенбіз. Соның нәтижссінде бұндай қосымшалардың екі түрлі тобы болатыны мәлім болды. Олардың біріншісі – зат есімдерге тіркелетін және басқа сөз таптарынан зат есімдер тудыратын жалғау-жұрнақтар, ал екінші топ – зат есімдерден етістік туғызатын жұрнақтар екен. Бұндай «Қосымшалар тіркесімінің сөздігін» сөздердің парадигматикалық эквиваленттілігін білу үшін де пайдалануға болады.

Сонымен, қазақ тілінде бір ғана зат есімдерге жалғанып, сөзді түрлендіретін қосымшалардың саны 500-ге таяу екеніне, ал әр алуан қазақ етістіктерінің 1000-ға жуық тұлғаны қабылдап, түрлене алатындығына біздің көзіміз жетті [4]. Нәтижесінде морфологиялық деректерді негізге ала отырып, қазақ тілінің сөз таптары мен грамматикалық категорияларын автоматты жолмен таптастыруды, сөйтіп оларды формализациялау (жігін айыру) қажеттігі келіп шықты. Бұл ретте біз мынаны ескердік: дәстүрлі түркі тіл білімінде сөз таптары негізінен әрбір лексикалық элементтің мағыналық мазмазмұнына, яғни басқаша айтқанда оның сөз тудыру мен сөз түрлендіру жүйелерінің қайсысына жататындығына және синтаксистік қызметіне қарай ажыратылады. Осы белгілері бойынша түркі тілдерінің негізгі және туынды морфемалары үш топқа бөлінеді. Олар:

1. Есімді, демек заттың атын немесе оның тұрақты белгісін білдіретін негізгі түбір морфемалар.

2. Етістіктен жасалған, іс-қимылды, яғни динамиканы білдіретін негізгі түбір морфемалар.

3. Сөйлемдер мен сөз тіркестерінің құрамында тұрып, лексика-семантикалық мағыналарды емес, тек өзара грамматикалық қатынастарды ғана білдіретін көмекші сөз ретіндегі түбір морфемалар.

Бұлардың алғашқы тобына зат есім, сын есім, үстеу, сан есім, одағайлар мен еліктеуіш сөздер, есімдіктер жатады; екінші топқа – естіктер, ал үшінші топқа – демеулік шылау, жалғаулық шылау, жалғаулықтар мен модаль мәнді сөздер жатады.

Қазақ тілінің табиғатын жыға тану үшін, сөз жоқ буын құрылымының жүйесін жете білу қажет. Қазақ жазуы көптеген үнді-европа тілдерінің жазулары секілді дыбысталу жүйесіне негізделген. Солай бола тұрса да, қазақ жазуында буын құрамының ықпалы басым. Сондықтан да қазақ тілінің морфологиялық құрылымын танып білу үшін қазақ сөздерін буын жігіне қарай сегменттеу, яғни ажырата білу шарт. Бұл орайда бізге үлкен көземел болған теориялық база деп профессор Қ.Жұбановтың 1934 жылы жариялаған «Буын жігін қалай табуға болады?» - деген еңбегін атауға әбден болады. Ғалым бұнда тіліміздегі буындардың негізгі алты белгісін басшылыққа ала отырып, қазақ сөздеріндегі буындардың жігін айырудың өзіндік алгоритмін жасаған. Төменде сол алты қасиеттердің сипаттамасын береміз:

- қазақ тілінде дауысты дыбыстың қатынасынсыз буын жасалмайды;
- бір буын ішінде біреуден артық дауысты дыбыс болмайды;
- дауысты дыбыс жеке-дара тұрып та буын жасайды;
- бір буынның құрамында үштен артық дауыссыз дыбыс болмайды;
- бір буынның ішінде қатар келетін екі дауыссыз дыбыс тек дауысты дыбыстан соң ғана тұрады және олар міндетті түрде: сонор, яки үнді+үнсіз дыбыс тіркесі түрінде ұшырасады.
- сөз ортасында дауысты дыбыстан басталатын буын болмайды. Сөз ортасында және сөз соңында тұратын буындар міндетті түрде инициал (әлфләм) дыбыстан басталады.

Қ.Жұбанов сөздің буын жігін ажыратуды (сегментация) сөздің ең соңында тұрған буыннан бастауды ұсынған. Бұл жағдайда сөздің соңынан бастап санағанда ұшырасатын әрбір дауысты дыбысқа оның алдында тұрған дауыссыз дыбыс қосылып барып жіктелуі шарт. Өйткені ол (дауыссыз) буынның жігін нұсқайтын сегмент болып табылады. Қазақ тілінің өзіне тән төл сөздерінде бір буынның құрамына енетін дыбыстардың саны төртеуден аспайды деп көрсетеді ғалым.

Дыбыстың құрамы әр алуан болып келетін қазақ буындарының алты вариантын профессор Қ.Жұбанов мынадай шартты белгілермен таңбалаған, қараңыз: ***Д-ты; Д-сыз+Д-ты; Д-ты+Д-сыз; Д-сыз+Д-ты+Д-сыз; Д-сыз+Д-ты+Д-сыз+Д-сыз***.

Қазақ тілінің белгілі бір сөз тұлғасына автоматты түрде морфологиялық талдау жасағанда, атап айтқанда текст ішіндегі кез-келген сөз тұлғасын өзімізге қажетті реестр бойынша тәртіптегенде біз қазақ сөздерінің осы сипаттамада аталған буындау жүйесіне сүйендік. Бұл мақсаттағы алторитмнің шешімі нақты бір сөздің құрамындағы буын жүйесінің ішінен сөз тудырушы және сөз түрлендіруші аффикстерді тауып,

іріктеу жұмыстарына септігін тигізуі тиіс. Сөздің соңғы буынынан басталып, ең ақырында сөз басындағы алғашқы буынды жіктеумен тынатын бұндай ретроспективалы талдау үстінде буын құрылымдары өзімізге қажетті, әрі күнілгері белгіленіп қойылған аффикстер тізбегімен салыстырыла отырады да, соңғы позициялары сөз түрлендіруші қосымшаларды бөліп тастауға мүмкіндік берді [5, 66]. Бұның өзі аффикстер жүйесіндегі омонимикалық ұқсастықтардың басын ашуға да септігін тигізеді. Ал, біздің сегменттерге жіктеу тәсілімізге бойсынбайтын кейбір әредік құбылыстар болса, оларды дисплей экранына сәулелендіріп, ары қарайғы дәлдестірулерді көзбен шолып отырып жүргіздік. Бұл тәрізді жалпы ережеге бағынбайтын жайттар айырықша құбылыстар кестесінде көрсетіледі. Осы сөз болып отырған алгоритмді біз стилі және әралуан текстерді талдауға да пайдаландық [5, 148-156].

Әдебиеттер

1. *Жұбанов Х.К.* Выступление на съсзде учителей-ударников Казахстана, - Материалы ЦГА РК. фонд 81, опись 3, дело 504, листы 17-22.

2. *Жұбанов Қ.* Қазақ тілінің ғылыми курсы жөнінен лекциялар, 1-кесек. Қазақ тілінің фонетикасы (дыбыс жүйесі). «Политехникалы мектеп», № 7-8. 1933. Сілтеме: Құдайберген Жұбанов. Қазақ тілі жөніндегі зерттеулер. - Алматы, 1966. Көлемі 362 бет.

ҚОЛДАНБАЛЫ ТІЛ БІЛІМІ

Қазақ тілі. Энциклопедия. Алматы: Қазақстан Республикасы Білім, мәдениет және денсаулық министрлігі, Қазақстан даму институты, 1998 ж., 240-бет.

Қоғамдық қажеттікті қанағаттандыру мақсатындағы проблемалар мен әдіс-амалдардың жиынтығы. Қ.т.б-нің дәстүрлі және жаңаша мақсаттарға сай түсініктері бар. Дәстүрлі түсінікке – сөздіктер түзу, әліпби мен жазу жүйесін өңдеу (зерттеу), ана тілі мен шет тілдерін оқытудың лингвистикалық негізінен құру, бір тілден екінші тілге аудару, орфография мен емлеге, тіл мәдениеті және сөйлеу шешендігі мәселелері жатады. Ал жаңаша мақсаттағы түсінікке: машинамен аудару, ақпараттар тілі, терминология және оның ақапарат тілі, терминология және оның ақпаратты іздеудегі рөлі, автоматты индекстеу, құжаттардың рефераты мен аннотациясын алу, ақапарат жүйесінің лингвистикалық негізін құру, ақпаратты іздеу қажеттігі үшін тезаурус – сөздіктер түзу, тіл құрылымын анализдеу мен синтездеу жұмыстарын автоматтандыру және т. б. мәселелер жатады.

Қ.т.б-нің жаңа проблемалары дәстүрлі түсініктегі мәселелерді терістемейді, тек қана тілшілер мен математиктердің тығыз одағы мен олардың ЭЕМ-мен қатынаста болуын қалайды. Дәстүрлі Қ.т.б-нің ең құнды проблемаларына жалпы және терминологиялық тіл мәдениеті жатады. Осыған байланысты тілдің коммуникативтік қасиетіндегі сапасын

анықтайтын сөздер мен ұғымдары терминдеу қажеттігі туады. Осы тұрғыдан алғанда терминдеуге жататындар: «дұрыстылық», «байлық», «дәлдік», «тазалық», «көркемділік», орындылық», «әрекеттілік», «нанымдылық», «қисындылық», «жеткіліктік». Қазақстанда қазақ тіліне қатысты Қ.т.б-нің дәстүрлі мәселелерімен қатар жаңадан туған ЭЕМ-ге байланысты мәселелері де өзінің шешімін табуда (қ. *Математикалық тіл білімі, Машиналық аударма, Машиналық қор, Лингвистикалық статистика, Жиілік сөздік).*

Әдеб.: Березин Ф.М., Головин Б.Н., Общее языкознание. М., 1979; Звегинцев В.А., Теоретическая и прикладная лингвистика. М., 1963.

МАШИНАЛЫҚ АУДАРМА

Қазақ тілі. Энциклопедия. Алматы: ҚР Білім,
мәдениет және денсаулық министрлігі, Қазақстан
даму институты, 1998 ж., 268-269 бб.

Автоматты аудару – аударма әдісін жеңілдету мақсатымен бір тілдегі тексті екінші тілге электрондық-есептеуіш машинаның (ЭЕМ) көмегімен аудару. М.а. инженерлі-лингвистика мен қолданбалы ғылымдар саласына жатады. Тарихи даму жағынан үш кезеңге белуге болады: 1946-1957 ж., 1957-1967 ж., 1967 жылдан осы кезеңге дейін. Бірінші кезеңге тән жағдай: М.а. жайлы идеяның тууы, зерттеушілердің осы мәселеге қатты қызығып, құлшынып кірісіп тез арада шешімін табуға бар ықыласын салуы. Екінші кезең – М.а. проблемасын шешудегі жұмыстардың ең бір жемісті кезеңі. Әсіресе осы кезге дейінгі тіл туралы деректердің жеткіліксіздігін, бұл проблеманың ірі теориялық мәселе екендігін мойындау кезеңі еді. Табиғи тілді код ретінде (шартты белгі) қарастырып, тілді өлшеуде статистикалық және теория-ақпараттық әдістер қолданыла бастады. Зерттеу салмағы көбінесе лексикадан синтаксис саласына ауысып семантикалық анализдің көмегімен текстің мағына жағына көп көңіл бөлінді. Көптеген тілдік мәліметтерді қамтитын кең көлемдегі машиналық сөздіктер түзілді. Осы жұмыстардың нәтижесінде өндірістің мұқтаждығына сай кейбір М.а. жүйелері іске қосылды. Үшінші кезеңде М.а. өзіне тән социалдық статус алды. Ғалымдар техник. кибернетиканың жаңа жетістіктерін автоматты түрде тілді зерттеудегі теория мен практиканың соңғы табыстарымен ұштастыруға, М.а-ның толық жүйесін құруға көп еңбек сіңірді. Осыған дәлел – 1977 ж. Люксембургте еткен «Преодоление языковых барьеров» атты конференция. М.а-ның мамандары табиғи тілді, біріншіден, ашық жүйелер тобына жататындығын және оның үнемі толығып отыратындығын, ал машина (ЭЕМ) үшін шекті жүйенің қажеттігін, екіншіден, табиғи тілмен ақпаратты баяндауда тілдің артықтық (избыточность) қасиетін де ескеру қажет ететіндігін айтты. Орташа есеппен алғанда тілдегі ондай артықтық 70-80 процентке дейін жететіні байқалды. Бұл шама әр түрлі функционалдық стильдерге қарай өзгеріп отырады. Мысалы ғыл.-техник. текстерде артық

хабар 85-95 процентке дейін өседі. Осы себептен М.а-ға ыңғайлысы ғыл.-техник. текстер болып саналады да, бір стильдің өзі ішкі салаларына қарай тағы да бөлшектенеді (подъязык). Соңғы бөлшектенген текстерден машиналық сөздіктер түзіліп және тілдің синтаксистік-семантикалық қасиеттері негізінде аударма алгоритмдері мен бағдарламалар жасалады. Осындай жұмыстардың нәтижесінде алынған М.а. сол күйінде тұтынушының қажеттілігіне жарамды бола қоймады, сондықтан ол аудармашы-мамандардың қолынан жеңіл-желпі өңдеуді қажет етеді. Бұл жұмыс әрине тексті машинаның көмегінсіз аударудан әлдеқайда жеңіл.

Қазақстанда М.а-мен Шымкент қ-ндағы М. О. Әуезов атынд. пед. ин-тында проф. Қ. Б. Бектаевтың жетекшілігімен шұғылданатын шағын ғыл. топ бар. Олар алғашында полимерлер химиясының патент текстерін зерттеп ағылшын тілінен орыс тіліне машинамен аудару ісін өндірістік мұқтаждыққа жарайтындай дәрежеге жеткізді. Соңғы кезде тәжірибе ретінде әр түрлі текстерді қазақшадан орысшаға аударатын М.а-ның мәселесімен шұғылдануда. Осы мақсатпен қажетті машиналық сөздіктер түзіліп, тиісті теориялық және практикалық зерттеулер жүргізілуде. Қазіргі кезде ЭЕМ-нің көмегімен іске қосылған ағылшын тілінен орыс тіліне аударатын – АМПАР, немісшеден орысшаға аударатын – НЕРПА, французшадан орысшаға аударатын – ФРАП сияқты аударма жүйелері бар.

Әдеб.: Аханов К., Тіл білімінің негіздері. А, 1973; Машинный перевод. М., 1957; Мельчук И. А, Равич Р. Д., Автоматический перевод. 1949-1963, М., 1967; Автоматический перевод. М., 1973; Пиотровский Р. Г., Отраслевой вероятностный машинный перевод. В сб.: «Статистика текста». Т. II., Минск, 1970; Бектаев К. Б., Джубанов А., и др. Опыт реализации микрословарей для машинного перевода английских патентов по химии полимеров. Всесоюз. конф. по маш. переводу, Дагестанский ун-т, 1978; Марчук Ю. Н., Проблемы машинного перевода. М., 1983.

ПРОФ. ҚҰДАЙБЕРГЕН ЖҰБАНОВТЫҢ БУЫН ТЕОРИЯСЫ ЖӘНЕ ОНЫҢ КОМПЬЮТЕРЛІК ЛИНГВИСТИКАДАҒЫ ОРНЫ

Қазақстан республикасының Мәдениет, ақпарат
және қоғамдық келісім министрлігі Тілдерді дамыту
департаменті. «Мемлекеттік тіл: терминология, іс қағаздары
мен бұқаралық ақпарат құралдарының тілі» атты
«Қ. Жұбановтың 100 жылдығына арналған республикалық
ғылыми-тәжрибелік конференция» материалдары.
Астана – 1999, 298-305-бб.

Буын теориясы – фонетикадағы ең күрделі мәселелердің бірі екені фонетист-ғалымдарға белгілі.

Проф. Қ. Жұбанов жазған «Қазақ тілі грамматикасының» фонетика бөлімінде дыбыс, дыбыстың саны мен сапасы, дауысты-дауыссыздар, олардың артикуляциясы мен акустикасы, дыбыстау аппараты, буын

сингармонизм, екпін жайлы жүйелі түрде сөз болады. Сондықтан да, ғалымдардың қазақ тілінің осы мәселелерін ғылыми тұрғыдан зерттеу мәселесін тұңғыш қолға алған ғалым – проф. Қ. Жұбанов деп жүргендері бекер болмасқа керек.

Тіл мамандары арасында көбінде, Қ. Жұбанов теоретик ғалым ба, әлде практик ғалым ба? – деген сұрақтарды жиі естуге болады. Осы жөнінен филология ғылымдарының докторы, Қазақстан Ғылым академиясының корреспондент мүшесі, проф. Рабиға Ғалиқызы Сыздықова былай дейді: «Теоретик ғалым деп айтуға негіз болатын тұжырымдардың бірі – ол қазіргі кезде көбірек назар аударылып жүрген «Валенттілік теориясы», «Текст теориясы», «Функционалдық грамматика», «Лингвистикалық статистика» деген тілдік бағыттар, дәл осы терминдермен аталмаса да, ғалымның ізденістерінен бұл теориялық мәселелердің алғашқы нысандарын көреміз»[1]. Айтқан сөздері дәлелді бояу үшін, Р.Ғ. Сыздықова, ғалым еңбектерінен тиісті үзінділерін келтіре отыра, басқа зерттеуші-ғалымдардың да осы жөніндегі пікірлерін қоса айта кетеді.

Фонологиялық бірліктердің негізінде тілдерді жіктеу кезінде түркі тілдерін, оның ішінде қазақ тілін, фонетикалық тіл, не буын тіл қатарына жатқызу керек пе? – деген сұраққа экспериментті-фонетика саласының белгілі маманы проф. Ә. Жүнісбеков: «Біздің байқауымызша, бұл мәселе тіл мамандарын осы күнге дейін толғандыра қоймаған тәрізді. Соның нәтижесінде түркі тілдері фонемалық тілдердің қатарына ат үсті жатқызыла салынған... Қазақ тілі дыбыстарының буын ішінде үндесіп барып сөз құрайтынын аңғарған ғалым «қазақ тілі – буыншыл тіл» деп соны жаңалық айтқан. Қазақ сөзін сөз етіп тұрған сингармонизм құбылысы екенін, ал сингармонизмнің иек артар тілдік бөлшегі буын екенін ескерсек, проф. Қ.Жұбановтың қаншалықты көрегендік танытқанын көреміз»[2].

Шынында да, Қ. Жұбановтың кең көлемде алып, жан-жақты зерттеген мәселелерінің бірі – қазақ сөздерінің буындық құрамы. Буынның тек дауысты дыбыстар арқылы ғана жасалатынын түсіндіре келіп, құрамындағы дыбыстар санына қарай бір дыбысты, екі дыбысты, үш дыбысты, төрт дыбысты деп бөліп әрқайсысының қандай дыбыстық құрамда болатындығына тоқталады. Ғалым буындарды оның құрамындағы дауысты дыбыстың тұрған орнына қарай жіктейді. Мысалы, буын бір дауыстыдан, немесе дауыссыздан басталып, дауыстыға аяқталса оны ашық буын деп атайды. Сол сияқты, дыбыстыдан басталып бір не екі дауыссызға аяқталса екі дыбыстысын – жеңіл тұйық, үш дыбыстысын – ауыр тұйық буынға жатқызады (*ер, ерт*). Буын құрайтын дауысты екінші орында келіп, одан кейін бір не екі дауыссыз келсе үш дыбыстысы – жеңіл бітеу, төрт дыбыстысы – ауыр бітеу делінген (*кан, кант*).

---

[1] *Сыздықова Р.Ғ.* Қ.Жұбановтың лингвистикалық мұрасын танып-игеру проблемалары // Жұбанов тағылымы. – Актөбе, 1998. 13-18-б.

[2] *Жүнісбеков Ә.* Қазақ фонетикасының ғылыми курсын жасаушы. «Құдайберген Жұбанов және қазақ совет тіл білімі». «Ғылым», Алматы, 1990, 193-194-б.б.

Буынның жылысуы, тасымал мәселелері де сөз болады. Проф. Қ. Жұбановтың «Буын жігін қалай табуға болады?»[3] атты мақаласының, біздіңше, әрі ғылымдық, әрі практикалық жағынан өте мәнді еңбек дейтініміз – бұл тек өз кезеңі үшін емес, бүгінгі, тіпті ертеңгі күн үшін де маңызы зор екендігі. Ғалым қазақ тілінің табиғатын дұрыс тану үшін буын жүйесін жете білу керектігін ескертіп: «...буынды жақсы білмесең, емлені де білмейсің; сөз тудыру, сөз түрлендіру сияқты морфология негізін де ұға алмайсың; буынды білмейінше, сөз екпіні (ударение), сөз сазы (интонация) секілді сөйлем жүйесіне қатысты тіл құбылыстарының сырына да қана алмайсың...», - деп буын жүйесінің құрылымын, оның алуан тілдердегі атқаратын қызметтерін атап көрсетеді.

Енді, хабарламамыздың екінші бөлігі – буын теориясының компьютерлік лингвистикадағы орнына қысқаша тоқталайық. Тілдің ішкі заңдылықтарын ашуға арналған бұрынғы, әсіресе, қазіргі зерттеулерде ғалымдар сандық тәсілді кеңінен қолданып, тілдік элементтердің (құбылыстардың) саны мен сапа байланыстарын анықтауға тырысқан. Тіл ғылымында сандық зерттеулердің қажеттігін тек математикалық лингвистиканы жақтаушылар ғана емес, олармен қатар дәстүрлі тіл білімі саласының белгілі өкілдері де мойындап жүр. Басқа ғалымдардың зерттеулеріне тоқталмай-ақ, қазақ тілінің тұңғыш проф. Қ.Жұбановқа қатысты жерлерін ғана айта кетпекпіз. «Қазақ тіл біліміндегі лингвистикалық статистика саласының бастауында да проф. Қ. Жұбанов тұрды... Осы бағытта қазақ тіл біліміндегі көптеген мәселелердің «Әліппесін» ғалым өзі бастап, өзі жазып кеткен болатын. Осылайша, біз сөз етіп отырған тілді зерттеудегі сандық мәліметтер мен деректерді пайдалану әдісі де аты аталып, түсі түстеліп, арнайы түрде қарастырылмағанмен, ол жөнінде жанама болса да сөз етіліп, белгілі бір игі мақсаттың жолын дәл мегзеп берген еді»[4], – дейді филология ғылымдарының докторы, математика кафедрасының профессоры Қ. Б. Бектаев. Осы мақалада проф. Қ.Бектаев ғалым Қ.Жұбановтың зерттеулеріндегі сандық тәсілге жүгініп, олардың сапаға әкеліп соғатын жерлерін түгелдей атап (бір оқығанда тілші-ғалымдар ескере қоймайтын) қызықты жәйттерді сөз ете келе, соңында: «Қ. Жұбанов сандық деректердің негізінде тілге қатысты бірсыпыра жайларды (дыбыстар, буындар, сөздер, сөйлем мүшелерінің орны т.б.) нақты анықтап, кейбір тың тұжырымдар жасайды. Мысалы, ғалым былай дейді: «Ілиястың не бәрі 12 жол «Қойшы ойы» деген кішкене өлеңінде 488 дыбыс бар. Бірақ, 488 дыбыс 488 түрлі емес, мұнда не бары 28-ақ түрлі дыбыс бар. Осы 28 түрлі дыбыстың бірі 10, бірі 50 рет қайталап барып 488 болған». Осындай есеп-қисаптан кейін, мынандай тұжырым жасаған: «Қазақ тілінде де сапасы әр түрлі болатын дыбыстардың арнаулы саны бар»[5].

---

[3] Жұбанов Қ. Қазақ тілі жөніндегі зерттеулер. «Ғылым», Алматы, 1966, 330-340-бб.

[4] Бектаев Қ.Б. Қазақ тіл біліміндегі лингвистикалық статистиканың алғашқы көріністері // Құдайберген Жұбанов және қазақ совет тіл білімі. «Ғылым», Алматы, 1990, 109-110-бб.

[5] Жұбанов Қ. Қазақ тілі жөніндегі зерттеулер. «Ғылым», Алматы, 1966, 169-170-б.б.

[6] Караулов Ю.Н. Методология лингвистического иследования и машинный фонд

Сандық тәсілді мейлінше кеңінен қолдану қажеттігінен, соңғы кезде, электронды-есептеуіш машиналардың (ЭЕМ) көмегін пайдалану керек болды. Үнді-европа тілдерін зерттеу барысында ЭЕМ-дер, ең алдымен: машиналық аударма, текстерді автоматты түрде индекстеу, аннотациялау, рефераттау, сөйлеу тілін анализдеу мен синтездеу және т. б. зерттеулерді іске асыру үшін қажет болғаны белгілі. Осы зерттеулердің негізінде математикалық лингвистика, құрылымдық лингвистика, инженерлік лингвистика, статистикалық лингвистика, ал соңғы кезде компьютерлік лингвистика деп аталатын тіл зерттеудің жаңа салалары пайда болды. Бұрынғы дәстүрлі әдіспен тіл зерттеу кезіндегі қиындыққа соқтыратын көптеген процестерді автоматтандыратын алгоритмдік тәсілдер кең түрде қолданыла бастады. Осындай өзі күрделі әрі көп уақыт қажет ететін жұмыстың түрлерін компьютерлік лингвистика өз міндетіне алды.

Түркі тілдерін, оның ішінде қазақ тілін жаңа тәсілмен зерттеуде де проф. Қ. Бектаевтың жетекшілік етуімен Қазақстандық «Статистика тобы» біршама жетістіктерге жетті деуге болады. Ал, Қазақстанның Ұлттық Ғылым академиясының А. Байтұрсынұлы атындағы Тіл білімі институтының «Қолданбалы лингвистика» бөлімінің ғылыми қызметкерлері жоспарлы түрде «Қазақ тілінің компьютерлік қорын» жасаумен көптен бері айналысуда. Осындай қомақты да және өте күрделі жұмыстың барысында, бізге қазақ сөз тұлғаларының құрамын статистикалық әдіспен зерттеп, формалды жақтарын айырып, оны автоматты түрде бөлшектеу (сегментация) ісімен шұғылдануға тура келді. Себебі, тілдің компьютерлік қоры деген ұғымның түпкі төркіні лексикографиялық, яғни сөздік түріндегі жұмыстардың идеясына келіп тіреледі. Осыған сәйкес, орыс тілінің машиналық қорын жасаушылардың бірі Ю.К.Караулов компьютерлік қордағы тілдік информацияларды шоғырландыратын бірегей түйін немесе түп қазық жеке сөз деп есептейді[6]. Өйткені, оның тұжырымдауынша, сөз тілдің ең табиғи бөлшегі болып табылады. Ол морфемалармен, сөз орамдарымен немесе сөйлемдермен салыстырғанда да ішкі-сыртқы қызметтері жағынан, яғни морфемаға, буынға, әріпке, мағынаға талданатындығы, сондай-ақ, сөздерді өзгертетіндігі, жаңа сөз тудыратындығы, синтаксистік, грамматикалық ассоциациялық және концептуалды қатарлар жасайтындығы жағынан ерекшеленеді.

Осы айтылғандарды басшылыққа ала отырып, алғашқы зерттеу ретінде қазақ текстері бойынша сөздердің статистикалық сипаттамасын жасауға көп күш жұмсадық. Алынған деректердің негізінде зерттеуші-тілшілердің ми қызметінің жұмысын, оның білімінің арқасында тапқыштық қабілетін формалдау проблемасымен шұғылдандық. Мысалы, айталық тілші сөйлемде кездесетін әр сөзді сөз таптарына, сөйлем мүшелеріне қалай ажыратады? Оларға ажырату кезінде формалды белгілердің (признактардың) көмегі тие ме? Не болмаса негізгі, туынды түбірлерді формалды белгілер арқылы жігін

русского языка//Машинный фонд русского языка: идеи и суждения. М., 1986, с. 13-25.

айыруға бола ма? - деген сияқты сұрақтарға жауап іздеу қажет болды. Яғни, адамның саналы түрдегі ми қызметін автоматты жолмен (санасыз) айырып тану мүмкіндігін арнайы зерттеу қажеттігі туды.

Осы және басқа да сұрақтарға жауап беру үшін, біз ми қызметінің жұмысын алгоритмдер тіліне көшіру арқылы іске асыруға тырыстық. Өз алгоритмімізде тек зерттелетін текстегі сөз тұлғалары мен олардың негізгі түбірлерін анықтап қана қоймай, сонымен бірге объектідегі сөз тұлғаларының құрамында сөзді тудырушы және сөзді түрлендіруші қанша қосымшалар тіркесі барын есептейтін «Сөздік» жасауды да мақсат еткенбіз. Соның нәтижесінде мұндай қосымшалардың екі түрлі тобы болатыны мәлім болды. Олардың біріншісі – зат есімдерді тудыратын жалғау-жұрнақтар екен де, ал екінші топ – зат есімдерден етістік туғызатын жұрнақтар екен.

Бұндай «Қосымшалар тіркесімінің сөздігін» сөздердің парадигматикалық эквиваленттілігін білу үшін де пайдалануға болатыны анықталды. Сонымен, қазақ тілінде бір ғана зат есімдерге жалғанып, сөзді түрлендіретін қосымшалардың саны 500-ге таяу екеніне, ал әр алуан қазақ етістіктерінің 1000-ға жуық тұлғаны қабылдап, түрлене алатындығына біздің көзіміз жетті[7]. Осы және басқа да деректер қазақ тілінің сөз таптары мен грамматикалық категорияларын формалды түрде топтастыруға өз үлесін қоспақшы. Бұл ретте біз мынаны ескердік: дәстүрлі түркі тіл білімінде сөз таптары негізінен әрбір лексикалық элементінің мағыналық мазмұнына, яғни дыбыстау кезінде оның сөз тудыру мен сөз түрлендіру жүйелерінің қайсысына жататындығына және синтаксистік қызметіне қарай ажыратылады. Осы белгілері бойынша түркі тілдерінің негізгі және туынды морфемалары үш топқа бөлінеді. Бұлардың алғашқы тобына зат есім, сын есім, үстеу, сан есім, одағайлар мен еліктеуіш сөздер, есімдіктер жатады; екінші топқа – етістіктер, ал үшінші топқа – демеулік шылау, жалғаулық шылау, жалғаулықтар мен модаль мәнді сөздер жатады.

Проф. Қ. Жұбанов сөздің буын жігін ажыратуды сөздің ең соңында тұрған буыннан бастауды ұсынған. Бұл жағдайда сөздің соңынан бастап санағанда ұшырасатын әр дауысты дыбысқа оның алдында тұрған дауыссыз дыбыс қосылып барып жіктелу шарт. Өйткені ол (дауыссыз) буынның жігін нұсқайтын сегмент болып табылады. Қазақ тілінің өзіне тән төл сөздерінде бір буынның құрамына енетін дыбыстардың саны төртеуден аспайды деп көрсетеді ғалым.

Қазақ тілінің белгілі бір сөз тұлғасына автоматты түрде морфологиялық талдау жасағанда, атап айтқанда текст ішіндегі кез-келген сөз тұлғасын өзімізге қажетті реестр бойынша тәртіптегенде біз қазақ сөздерінің осы сипаттамада аталған буындау жүйесіне сүйендік. Бұл мақсаттағы алгоритмнің шешімі нақты бір сөздің құрамындағы буын жүйесінің ішінен сөз тудырушы және сөз түрлендіруші аффикстерді тауып, іріктеу жұмыстарына септігін тигізуі тиіс. Сөздің соңғы буынынан басталып, ең ақырында сөз басындағы алғашқы буынды жіктеумен тынатын бұндай

[7] Қазақ тексінің статистикасы. Алма-Ата, 1973. 39-б.

ретроспективалы талдау үстінде буын құрылымдары өзімізге қажетті, әрі күнілгері белгіленіп қойылған аффикстер тізбегімен салыстырыла отырады да, соңы позициядағы сөз түрлендіруші қосымшаларды бөліп тастауға мүмкіндік береді. Мұның өзі аффикстер жүйесіндегі омонимиялық ұқсастықтардың басын ашуға да септігін тигізді. Ал біздің сегменттерге жіктеу тәсілімізге бойсынбайтын кейбір әредік құбылыстар болса, оларды дисплей экранына сәулелендіріп, ары қарайғы дәлдестірулерді көзбен шолып отырып жүргіздік. Бұл тәрізді жалпы ережеге бағынбайтын жайттар айырықша құбылыстар кестесінде көрсетіледі.

Қазіргі компьютерлердегі текстік редакторлар қазақ тексін компьютер жадына енгізуге, оны автоматты түрде, яғни адамның қатынасынсыз жөндеуге, қазақ сөздерін автоматты түрде тасымалдауға бейімделмеген. Сондықтан, қазақ текстерімен компьютерде жұмыс істейтін әр мекеме, не шағын топтар қазақ алфавитіне ғана тән әріптерді өзінше белгілеп (кодтап), өзінше бағдарламалар жасап жағдайларын құтқаруда. Мұның өзі біріздiлікке жол бермеушілік болып шығады. Мысалы, біз де өз жұмысымызда баяндалған мәселелердің шешімін табуда жоғарыда сөз болған проф. Қ. Жұбанов ұсынған сөзтұлғаны буынға бөлу теориясына жүгіне отырып қиындықтарды жеңуді жөн көрдік. Осындай біріздiліктің жоқтығы, бір мекемеде, немесе бір қалада дайындалған казақ тіліндегі ақпараттар екінші мекеменің, не басқа қаланың компьютерінде оқылмай қиындыққа ұшырататыны жиі кездеседі. Осыған орай, бірінші ұсыныс ретінде айтатыным, көпізділікке жол бермеу үшін, компьютерлік «қазақша драйвер» жайында мемлекеттік стандарт қабылдау қажет.

Сөз соңында айтпақшы екінші ұсынысым, ол – алғашында «Қазақ тілінің компьютерлік қорын» жасау мәселесін, болашақта салыстырмалы «Түркі тілдерінің компьютерлік қорын» жасау проблемасын Мемлекеттік іске айландырып, осындай өте маңызды және болашағы зор жұмысты, бір орталыққа бағындыру жағы ойластырылса.

II БӨЛІМ

ҚАЗАҚ МӘТІНІНІҢ МАЗМҰНЫНЫН ФОРМАЛДАУ

***

ФОРМАЛИЗАЦИИ СОДЕРЖАНИЯ КАЗАХСКОГО ТЕКСТА

СТАТИСТИЧЕСКОЕ ИССЛЕДОВАНИЕ КАЗАХСКОГО ТЕКСТА НА УРОВНЕ АБЗАЦЕВ

ҚР БжҒМ А.Байтұрсынұлы атындағы
Тіл білімі институтының «Тілтаным»
журналы. 2001. № 2. С. 35-41.

Текст, как объект текстологии и стилистики, привлекал внимание ученых и раньше. Текстологи, в основном, занимались литературно-филологическим изучением древних текстов. А стилисты, с целыо изучения стилистических (экспрессивных) эффектов, достигаемых в результате искусного использования авторами лексических и грамматических элементов языка. Объектом своего исследования они рассматривали, как правило, художественные произведения.

Языковеды прошлого также рассматривали объектом своего лингвистического изучения готовые тексты. При этом изучение текста было направлено прежде всего на анализ данного языкового материала. Однако, вопрос о том, как возникли эти тексты, оказывался за пределами интереса языковедов. Он мог интересовать текстологов или историков литературы, но с другой, нелингвистической точки зрения.

В этой связи необходимо отметить и заслуги представителей системно-структурной парадигмы, которые занимаясь описанием все более высоких уровней языковой системы, создали базу для изучения нового объекта – текста в его статике и динамике. Но эти исследования текста сводились в основном к внутриязыковому анализу, описание семантики языковых единиц ограничивалось пределами языка, без учета экстралингвистических

факторов. Семантическая теория описывалась вне связи со знанием о мире его носителей, тем самым по существу без учета человеческого фактора. По этой причине возникла необходимость создания адекватной теории значения, которая эксплицировала бы способность человека понимать данный язык и использовать его осмысленно в коммуникации.

Общепризнанным на сегодня стал тот факт, что естественный язык по существу есть не что иное как инструмент, позволяющий индивидиуму строить во внешнем мире знаковые модели, более или менее адекватно объективирующие фрагменты его концептуальной системы. Действительно, как пишет Луи Ельмслев, «Язык организован так, что с помощью горстки фигур и благодаря их все новым и новым расположениям может быть построен легион знаков». Многие ученые-лингвисты такие, как А.И.Смирницкий, В.М.Солнцев и др., поддерживая точку зрения Ф.де Соссюра, считая, что слово есть «нечто центральное во всем механизме языка», признают за «словом» статус основной языковой единицы. При этом за основные требования к единицам языка В.М.Солнцев предлагает считать следующие, т.е. в процессе функционирования языка его единицы должны:

а) выражать некоторый смысл либо участвовать в его выражении и дифференциации;

б) быть выделимы или вычленимы в качестве некоторых объектов;

в) быть воспроизводимы в готовом виде и др.

Определение понятия «текст» приведено во многих научных работах, но все они, в основном, рассматривают текст как некое структурированное по определенным законам единство, состоящих из языковых единиц – предложений, объединенных между собой и образующих более крупные единицы – сверхфразовые единства, абзацы, параграфы, главы, разделы и т.д., служащие для передачи определенного законченного содержания, отвечающего целям коммуникации.

С точки зрения лингвистики текста необходимо различение содержания языкового и смыслового, понятийного, обобщенного содержания текста. А.В.Бондарко пишет: «... план содержания текста структурирован лингвистически ,.., а смысл текста структурирован понятийно». Таким образом, конкретное содержание какого-либо текста членится на составляющие его элементы, на части, главы, на отдельные отрезки текста и на сверфразовое единство (СФЕ). В свою очередь основное содержание СФЕ состоит из составляющих его ведущих предложений. При этом смысл текста не является совокупностью смысловых образующих его предложений. Каждое предложение несет с собой «невидимый» (т.е. лингвистически не выражаемый) подтекст.

Из сказанного следует, что исследование лингвистической структуры текста необходимо начинать с выявления основных единиц, из которых строится текст. При этом основное внимание необходимо обращать на языковые и коммуникативные структуры предложения, а также и на другие компоненты, превышающих предложение и, наконец, – текста в целом.

С провозглашением казахского языка как государственного все более будет расти его роль в управлении народным хозяйством, в обучении и воспитании, в

культурной жизни и информационном обеспечении суверенной республики. Эффективное развитие всех упомянутых областей невозможно без создания автоматизированной системы управления народным хозяйством республики. Естественно, что такие системы должны функционировать на базе казахского языка. Значительно могут ускорить процесс активного внедрения в практику казахского языка создание автоматизированных систем обучения казахскому языку в школах и вузах республики. Не секрет, что как новое средство обучения, компьютер вызывает большой интерес у обучающихся, дает значительный эффект при обучении.

Одна из основных причин слабого использования компьютеров в информационной технологии на базе казахского языка и в обучении казахскому языку состоит в отсутствии формальных описаний казахского языка, которые могла бы воспринять электронная машина. Существующие грамматики казахского языка создавались в свое время для использования человеком. Правила таких грамматик не форматируемые или плохо форматируемые, так как в основном обращены к интуиции носителя языка, его лингвистическим и экстралингвистическим знаниям. Поэтому необходимо создавать описание языков, в том числе и казахского, которые были бы понятны компьютерам. Создание подобных описаний наталкивается на ряд особенностей, специфичных для казахского языка. К числу таких особенностей относятся сложность, неоднородность, недетерминированность и нечеткость языка. Недетерминированность казахского языка заключается в нечеткости областей значений слов и выражений, в нечеткости границ синтаксической правильности выражений этого языка. Эти особенности казахского языка и создают значительные трудности при построении полного описания казахского языка для его использования в системах автоматического анализа и синтеза казахского текста, в автоматизированных системах обучения языку и обучения гуманитарным дисциплинам на базе казахского языка.

Цель нашего исследования состоит в формализации содержания казахского текста, как основной единицы обучения. Такая проблема изучения казахского языка, можно сказать, ставится на повестку дня впервые.

Определение языка как «важнейшего средства человеческого общения» относится к более общим определениям. Поэтому, учитывая сложность образования языка, В.М.Солнцев вводит конкретизацию в определении, в зависимости от того с какой точки зрения может быть он рассмотрен. В книге «Язык как системно-структурное образование» автор приводит пять таких определений языка. Он пишет, что «... с семиотической точки зрения – язык есть система знаков, т.е. система материальных предметов (звуков), наделенных свойством обозначать что-то, существующих вне их самих, а также с точки зрения теории информации – язык есть код, с помощью которого кодируется семантическая информация...».

Новая область языкознания – лингвистика текста, возникшая на основании идеи необходимости принципиального различения языка и речи, рассматривает текст уже под другим углом зрения. По этому поводу Е.А.Реферовская пишет: «... текст не является простой фиксацией речи во всех ее формах, а обладает определенными законами построения, смыслового и

формального соединения и разграничения составляющих его единиц. ... Предмет лингвистики текста изучения – построение, порождение текста, его конструктивные элементы с точки зрения как формальной, так и содержательной структуры формирования его смысла».

При изучении законов порождения текста, т.е. когда язык изучается с помощью синтеза, выявляются те языковые закономерности, которые действуют при объединении слов в предложения, предложений в абзацы и т.д. Знание таких закономерностей поможет не только поновому подойти к проблемам машинного перевода, автоматического реферирования, аннотирования и поиска текстов, но и создать процедуры синтеза различных стандартных документов.

Таким образом, подход к тексту с точки зрения его порождения позволяет раскрыть такие важные закономерности естественного языка, которые во многих случаях остаются не выясненными при других методах анализа.

Известно, что любой текст отражает определенный фрагмент объективной действительности. Автором текста эта действительность представляется как совокупность огромного числа взаимосвязанных между собой предметов, явлений, свойств, процессов, отношений и т.д.

Пронизывающим весь текст во всей его протяженности и отражающим его смысловое единство являются, так называемые, «опорные слова». Эти слова в лингвистических источниках встречаются в различной терминологии: «ключевые слова», «смысловые вехи», «топики», «дескрипторы», «ядра», «фокальные точки» и т.д.

Дифференцированный анализ таких «опорных слов» показывает, что в его структуре существует определенная иерархия. А именно, часть таких «опорных слов» является «главными», «важными словами», «номенклатурными дескрипторами», «ключевыми словами», «выступающими точками» всего содержания текста. Можно сказать, что эти слова, определяя главный предмет сообщения, представляют из себя как бы свернутый замысл текста. Можно отметить и о наличии в тексте контекстуальных синонимов, т.е. таких слов, когда один и тот же референт может выражаться синонимическими словами и т.д.

Проблема выделения основного и факультативного содержания текста является одной из важных проблем инженерной лингвистики. В нашей работе предлагается два формальных критерия отделения главных опорных слов (ГОС), т.е. в его основе лежит идея о том, что такая наиболее значимая единица текста должна иметь, во-первых, наибольшую частоту в тексте, во-вторых, встречаться в наибольшем количестве абзацев такого текста.

Как показывают результаты наших и других подобных исследований, такими опорными словами, как правило, являются имена, отражающие главные действующие лица или главные предметы таких типов текста, как описание, повествование или рассуждение. Выделенные из анализируемого текста главные опорные слова в процессе развертывания содержания становятся своеобразным центром, вокруг которого формируются другие элементы, как составляющие микроситуации. Такими составляющими обычно служат такие классы слов как глаголы, прилагательные, наречия и др.

Известно, что в тексте переплетаются различные аспекты: структурно-коммуникативный и номинативный, социальный и индивидуальный,

статический и динамический, детерминированный и вероятностный и т.д. Только учитывая эти факторы можно раскрыть закономерности, которые руководят переходом некоторого «мысленного образования», содержания текста, в цепочки взаимосвязанных предложений, т.е. в текст.

Сложность казахского языка, также как и любых других, заключается в наличии в нем большого числа разнородных единиц (фонем, морфем, слов и т.д.), способных вступать в огромное число разнообразных отношений. Суть неоднозначности языка в том, что при любом способе классификации этих единиц, число таких классов достаточно велико и разброс количества единиц, входящих в эти классы, также достаточно велик. Недетерминированность выражается в том, что отношения между формой единицы и ее содержанием (словом и его значением, предложением и его смыслом, текстом и его содержанием) многозначны. Наконец, недетерминированность казахского языка заключается в нечеткости областей значений слов и выражений, в нечеткости границ синтаксической правильности выражений этого языка.

Эти особенности казахского языка и создают значительные трудности при построении полного описания казахского языка для его использования в системах автоматического анализа и синтеза казахского текста, в автоматизированных системах обучения языку и обучения гуманитарным дисциплинам на базе казахского языка.

В процессе исследования, т.е. при изучении процедуры вероятностно-алгоритмического порождения казахского текста, нами была сформулирована основная идея и уточнены следующие гипотезы:

1. Текст является отражением некоторой жизненной ситуации, воспринимаемой с точки зрения конкретного автора. Однако сам процесс восприятия обусловлен социально, т.е. любой автор членит действительность, а с ней и текст, на социально значимые «куски». Минимальным смысловым «куском» или «квантом содержания» является абзац, показывающий некоторую микроситуацию.

2. Основой или темой текста служит некоторое число опорных слов, которые отражают главные субъекты, объекты, основное место и время событий, описанных в тексте. Они являются своеобразными смысловыми вехами всего текста. Такие опорные слова можно объединить в таблицу основного статистического содержания текста (ТОСС).

3. Динамика текста понимается нами как «приведение в действие» его опорных слов, т.е. последовательная цепь их отношений между собой и с другими словами в разных частях текста. Наименьшей единицей динамической составляющей текста является семантико-синтаксическая формула абзаца (СЕСФА), отражающая указанные отношения между составляющими некоторой микроситуации внутри всей ситуации, описанной в тексте. Каждая СЕСФА связана последовательностью семантико-сиитаксических формул предложений (СЕСФП), записанных на специальном алгоритмическом языке (СЕМСИНТ).

4. Конкретизация отношений главных субъектов и объектов в пределах микроситуаций, а также уточнение прочих составляющих микроситуации, описанной в абзаце, осуществляется таблицей слов-конкретизаторов (ТАСК), которой снабжается каждая СЕСФА. Таблица слов-конкретизаторов (ТАСК) строится

на основе ролей, выполняемых словами абзацев в СЕСФП.

5. Текст с заданной таблицей основного статистического содержания (ТОСС) представляется в виде семантико-синтаксической формулы текста (СЕСФТ), состоящей из последовательности СЕСФА, типичных для данного автора или рассматриваемой предметной области.

6. Подбор СЕСФА для заданного содержания и порядок их следования в пределах СЕСФТ определятся двумя типами факторов: вероятностными, выявленными в процессе статистического анализа следования абзацев с разным предметно-логическим содержанием в текстах конкретной предметной области; детерминированными, полученными в результате качественного изучения взаимосвязи ТОСС, ТАСК и СЕСФА.

Главными при этом является детерминированные правила.

7. Лексическое наполнение каждой СЕСФА проводится также с проверкой двух типов условий: детерминированных, определяемых конкретным содержанием ТООС, ТАСК и однозначными кодами местоимений и служебных слов и логико-смысловых скрепов, а также связанных с вероятностным выбором прилагательных, причастий, наречий из списков слов, типичных для данной области знаний.

8. Грамматическое оформление каждого предложения, входящего в СЕСФА, осуществляется по строго детерминированным правилам, опирающимся на синтаксические значения семантических падежей, лежащих в основе языка СЕМСИНТ, и грамматические значения слов всех других классов. Видовременной аспект текста задается указанием при словах-конкретизаторах соответствующей грамматической информации.

Реальность этих гипотез доказывалась в результате проведения следующих процедур:

- статистического выделения всех зафиксированных в изучаемых текстах тем и микротем;
- выяснения способов логического объединения микротем, специфических для текстов определенного содержания;
- выделение инвариантных лингвистических единиц, специфичных для текстов каждого подъязыка.

Исходя из этого в работе последовательно строятся формальные модели распознания значения слов, абзацев и целого текста.

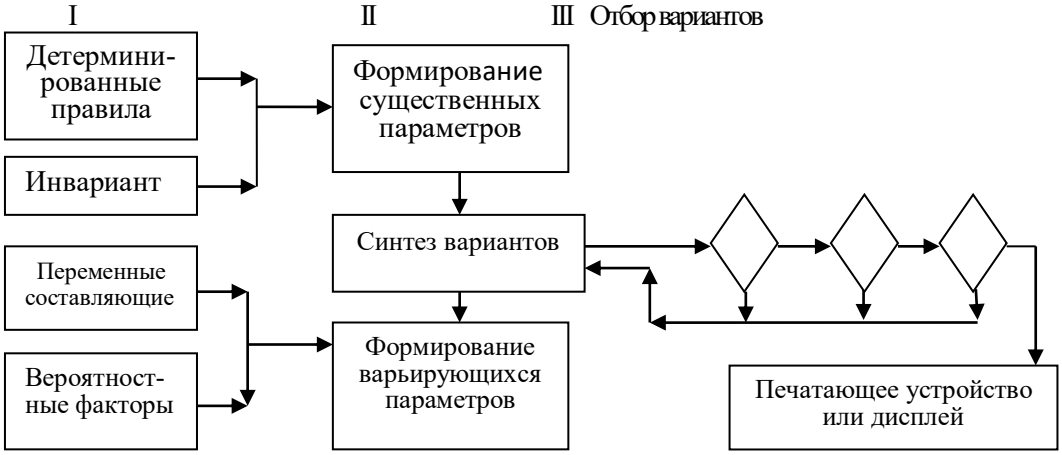
Реализация этих идей показана на схеме (см. рисунок).

Часть I этой схемы связана с выявлением инвариантов текстов определенной предметной области. Части II и III представляют процесс порождения текста. В части II выбираются очередные лингвистические единицы (СЕСФА, слова из ТОСС и ТАСК) с учетом как детерминированных правил, так и случайных факторов. Соответствие выбранных единиц определенным требованиям проверяется в части Ш. В итоге в памяти компьютера должен получиться некоторый текст (или его фрагмент), записанный на внутреннем языке компьютера. «Расшифровка» его осуществляется при выходе на печатающее устройство или на экран дисплея.

Для реализации этой схемы и построения детальных алгоритмов, выявления наиболее употребительных слов, СЕСФП, СЕСФА и ТАСК проводится анализ трех видов казахских письменных текстов – художественные, публицистические и

научно-популярные, характеризующиеся как описание, повествование и рассуждение к которым относятся: художественные поизведения (рассказы) известных казахских писателей, таких как С.Сейфуллин, Б.Майлин, М.Ауезов, С.Муканов, Г.Мусрепов, С.Ерубаев. Тексты шести рассказов указанных авторов состоят из 20 минитекстов, включающих в себя 500 абзацев.

Общая схема вероятностно-алгоритмического порождения текста



По научно-популярному жанру были исследованы тексты двух монографий (Калыбеков Т. Экология және ашық кен; Көшербаев К. Минералдар әлемінде). Число абзацев в указанных двух монографиях составляет также 500 абзацев.

Публицистические тексты были взяты из газеты «Қазақстан мұғалімі», вышедших с января 1986 года по январь 1987 года под рубрикой «Алты жасарлар мектепке барады», где обсуждаются проблемы новой школьной программы. Исследованию подвергнуты 35 газетных статей, охватывающих 102 абзаца. Таким образом, по текстам трех стилей были рассморены структуры 1500 абзацев.

Выбор именно этих текстов объясняется тем, что они представляют три наиболее употребительных функционально-смысловых типов речи (повествование, рассуждение, описание), без детального изучения которых невозможно построить универсальную модель распознавания содержания казахского текста.

Рассматривая проблему порождения текста в более широком плане, следует отметить ее особую значимость для теории и практики создания систем искусственного интеллекта. Опыт их функционирования показывает, что повышение эффективности таких систем возможно только путем использования для человеко-машинного общения естественного языка.

К ПРОБЛЕМЕ СЕМАНТИЧЕСКОЙ КЛАССИФИКАЦИИ ЗНАМЕНАТЕЛЬНЫХ КЛАССОВ СЛОВ КАЗАХСКОГО ЯЗЫКА

«Известия» МОН РК, НАН РК.
Серия филологическая.
Алматы-2001. № 2. С. 13-30.

В соотношении «язык – реальная действительность» язык служит не только обозначением реальной действительности, но в то же время является носителем обобщенного отражения действительности, органически соединяясь с соответствующими понятиями или отдельными их признаками, лежащими в основе семантических ценностей языковых единиц.

Общеизвестно, что лексические средства, наряду со звуковыми и грамматическими, составляют основу любого языка, в частности, и казахского. Слово, называя предметы, явления реального мира в их обобщенном виде, выступает в языке основным средством номинации. Являясь структурной ячейкой строя языка, оно имеет определенную морфологическую оформленность и грамматическую значимость. Слово охватывает фактически все языковые функции: номинативную (обозначения), сигнификативную (обобщения), коммуникативную (общения) и прагматическую (экспрессивного выражения и эмоционального воздействия).

Словарные знаки слов таких классов, как существительные, глаголы и прилагательные, обладают двумя основными семиологическими функциями – наименования и описания. Совместимость их в пределах слова порождает назывные высказывания, предикативные и лексические синтагмы. Они являются тем основным звеном в словарном составе языка, той постоянной значимой основой, которая обеспечивает «переход» от языка к мысли, от системы к речи, от слов к синтагматическим единицам, большим чем слова.

При семантической классификации лексики казахского языка имеются свои трудности и своя специфика. Это одна из труднейших задач семантики, которая не находит единого решения и в других языках мира. Количество таких классов и подклассов, их состав зависят от целей исследования, лингвистической школы и целого ряда других факторов.

Работ, посвященных исследованию казахского языка в области семантики, относительно мало. Здесь стоит назвать имя М.Оразова, его монографию «Қазақ, тілінің семантикасы» и его исследования по семантике казахского глагола [1; 2], которые затрагивают многие вопросы семантики казахского языка. Имеется и ряд других работ по некоторым проблемным вопросам семантики казахского языка [3]. Хотелось только отметить, что результаты этих работ во многом помогли нам при решении поставленных перед нами задач по вопросам формализации содержания казахского текста.

В самом общем виде можно выделить три принципа семантической классификации лексики казахского языка: 1) денотативно-десигнативный; 2) парадигматический; 3) синтагматический [1; 4; 5].

Денотативно-десигнативный принцип учитывает онтологическое расчленение действительности. При этом слова объединяются в группы в соответствии с тем, что они обозначают в реальной действительности (названия животных, названия растений, наименования определенных

классов предметов и т.п.). При парадигматической классификации слова объединяются в группы на основе тождества определенных семантических компонентов, выделенных в значении слов (модальные и каузативные глаголы и т.п.). По синтагматическому принципу слова объединяются в лексико-семантические группы с учетом количества и качества валентностей классифицируемых слов [1, 189-192; 4, 39-42].

Исходя из целей исследования за основу нашей семантической классификации принят денотативно-десигнативный принцип, хотя в некоторых случаях он будет дополняться двумя другими перечисленными выше принципами.

Семантическая классификация имен существительных (зат есім) казахского языка. Как известно, основным критерием разграничения имен существительных (наименований явлений, предметов, процессов, понятий) казахского языка, как и в других языках, на отдельные типы является прежде всего анализ их прямого номинативного значения, детерминированного типом денотата и сигнификата. При этом «слово» можно рассматривать двояко, т. е. как единицу первичного означивания в системе номинативных средств языка и как единицу, семантически реализованную в речи. Нами различаются две модификации «денотатов»: денотат как объект обозначения единичного предмета или класса предметов, а также денотат как предмет референции знака в речи.

А.А.Уфимцева в своей книге «Типы словесных знаков» [6, 102] подразделяет сигнификат как компонент знакового значения слов на шесть понятий, которые, как нам кажется, справедливы и для казахского языка. Имена с денотативным и денотативно-сигнификативным значением относятся к разряду имен, называемых конкретной лексикой, а слова с сигнификативной и сигнификативно-денотативной основой значения составляют абстрактную лексику любого языка, в том числе и казахского. Необходимо заметить, что определяющим для семантики именных лексем является обязательное наличие в их знаковом значении денотата и сигнификата. Если словесный знак служит специализированным названием (термином) в какой-либо узкой предметной области, то доминирует денотат над сигнификатом. В случае, когда словесный знак служит названием абстрактного понятия, то, наоборот, или сигнификат доминирует над денотатом, или словесный знак имеет чисто сигнификативное значение.

Именами существительными в казахском языке, так же как и во многих тюрских языках, являются различные наименования предметов и понятий в широком смысле слова. Это различные имена: собственные, например имена людей, географические названия и т.д.; нарицательные – названия человека: кісі *(человек), шеше (мать), әйел (женщина);* названия животных; названия растений; названия различных веществ и предметов быта; научные, технические и другие термины; наименования разных обобщенных отвлеченных понятий: *ел (народ), ру (род), тағдыр (судьба)* и т.д.; наименования действий, явлений, состояний: *жүріс (ходьба), хабар (известие), ұйқы (сон)* и т.д.; названия результатов действия: *жиналыс*

(собрание; ср. жинау – собираться), кеспе (лапша; ср. кесу – резать); названия орудий (инструментов) действия: *бұрғы (сверло; ср. бұрау – винтить, завинчивать), сыпырғыш (метла, веник), ара (пила)* и т.п.

Как пишет М.Оразов, среди слов казахских существительных самыми частыми по сложности их значений являются нарицательные имена. Эти имена существительные по отношению к значению имеют в языке изменчивый характер, т.е. иногда у них превалирует денотативное значение, иногда сигнификативное или же приобретается как бы суммарное значение – и динотативное, и сигнификативное. Примером может служить использование слов в метафорическом и метомическом значениях. Безусловно, когда слова употребляются в переносном значении, то изменяется и их денотативное и сигнификативное значение. Для убедительности сказанного можно сравнить слово «нар» в двух следующих предложениях: *Қырық балам бар еді, қырқы да бірдей нар еді.* (Батырлар жыры). *Айырдан туған жампоз бар – Нарға жүгін салгызбас* (Махамбет) [1, 98]. В первом предложении слово «нар» применено в переносном значении как сильный, «боевой», а во втором – в прямом значении как одногорбный верблюд. Их денотативные и сигнификативные значения несравнимы.

В количественном отношении тематические группы имен существительных в любом языке, в том числе и в казахском языке довольно велики. М.Оразов для примера приводит 8 разновидностей таких групп: 1) родственные термины (туысқандық, атаулар); 2) человек и названия частей его тела (дене мүше атаулары); 3) названия домашных животных (төрт түлік мал атаулары); 4) названия одежды, обуви человека (киім-кешек атаулары); 5) название птиц (құс атаулары); 6) название диких животных (жабайы аң атаулары); 7) название обиходных предметов-сосудов (посуды) (ыдыс-аяқ атаулары); 8) названия болезни человека (адамға тән ауру атаулары).

В нашем исследовании в основу семантической классификации казахских существительных положено деление реальной действительности на определенные зоны, сегменты [5 и др.]. С учетом целей нашего исследования, анализиуемого материала, а также отдельных идей, изложенных в работах [1; 3; 7] нами выделены семантические классы существительных (см. ***табл. 1***). Как видно из таблицы, денотативно-десигнативный принцип использован здесь при выделении наиболее крупных семантических классов существительных. Более мелкие лексико-семантические группы, например «части тела человека», «воинские звания», «дни недели» и др., связаны парадигматическими отношениями.

Семантическая классификация глаголов. Знаковое значение глагольных лексем. Имена существительные отражают и обозначают предметы или явления, а глаголы – отношения последних. Значение имен существительных – более автономно и абсолютно. Значение глагольных лексем – относительно и детерминировано многими факторами. В глагольных лексемах, выражающих понятие отношения, их взаимодействие как бы выносится за рамки глагольной лексемы и переносится в область синтагматических отношений действия к его объекту или субъекту (или

одновременно). Поэтому соотнесенность глагольных лексем с предметным рядом предполагает наличие смысловых связей глагола с предметными именами.

Таблица 1. **Семантические классы существительных, использованных в модели порождения казахского текста**

Сегменты реальной действитель-ности	**Наименование семантического класса**	**Код класса**
1. Наука	Общефилософские понятия	ЗТ01
	Общенаучные понятия	ЗТ02
	Общемедицинские понятия	ЗТ03
	Основные понятия психологии	ЗТ04
	Основные понятия нейропсихологии	ЗТ05
	Основные понятия логики	ЗТ06
	Основные понятия геометрии	ЗТ07
	Основные географические понятия	ЗТ08
	Основные понятия языкознания	ЗТ09
	Основные понятия литературоведения	ЗТ10
	Общемедицинские понятия	ЗТ11
	Основные понятия музыкального искусства	ЗТ12
	Основные юридические и уголовные понятия	ЗТ13
	Основные понятия логики	ЗТ14
	Основные понятия геометрии	ЗТ15
	Основные географические понятия	ЗТ16
	Основные понятия языкознания	ЗТ17
	Основные понятия литературоведения	ЗТ18
	Основные понятия изобразительного искусства	ЗТ19
	Основные понятия музыкального искусства	ЗТ20
	Основные юридические и уголовные понятия	ЗТ21
	Понятия, связанные с авиацией	ЗТ22
	Понятия, связан-е с устройством отдельн. Механ-го объекта	ЗТ23
	Информационные понятия	ЗТ24
2. Природа	Оценка	ЗТ31
	Предметы природы	ЗТ31
	Явления природы	ЗТ32
	Предметы природы – артефакты	ЗТ33
	Состояния природы	ЗТ34
	Понятия, описыв-ие структуру предм-в и части предметов	ЗТ35
	Вещества, металлы	ЗТ38
	Действия предметов природы	ЗТ39
3. Человек и его мир	Человек и его части тела	ЗТ40
	Части лица и головы человека	ЗТ41
	Состояния и ощущения, специфичные для человека	ЗТ42
	Действия, спецецифичные для человека	ЗТ43
	Понятия, используемые при характеристике человека	ЗТ44
	Предм-ты, одежды и обуви чел-ка и понятия, связан. с этим	ЗТ45

	Предметы-артефакты, используемые человеком в целях гигиены, для украшения и т.п.	ЗТ46
	Объекты, используемые для передвижения человека и понятия, с этим связанные	ЗТ47
	Предметы питания человека	ЗТ48
	Понятия, связанные с жилищем человека	ЗТ49
	Понятия, связанные со смертью и захоронением	ЗТ50
	Имена, отчества и фамилии людей	ЗТ51
	Национальность человека	ЗТ52
	Предметы-сосуды, которыми пользуется человек, и связанные с этим понятия	ЗТ53
	Предметы, используемые на отдыхе, охоте, рыбалке, и связанные с этим понятия	ЗТ54
	Наименование гражданских специальностей, должностей, степеней, званий	ЗТ55
	Предметы-артефакты, использ-мые при учении, труде и т.п.	ЗТ56
	Понятия, связанные с взаимоотношением людей	ЗТ57
	Понятия, связан. Родствен-ми отношениями между людьми	ЗТ58
4. Животный мир	Награды человека	ЗТ59
	Крупные домашние животные и части их тела	ЗТ60
	Насекомые и части их тела	ЗТ61
	Птицы и части их тела	ЗТ63
	Действия, спецефичные для животных	ЗТ64
	Предметы питания животных, птиц и насекомых	ЗТ65
5. Растения	Растения и их части	ЗТ66
	Деревья и их части	ЗТ67
6. Война и понятия, с ней связанные	Предметы боевой техники, оружия и их части	ЗТ68
	Наименование военных специональностей	ЗТ69
	Воинские звания	ЗТ70
	Понятия, связанные с оказанием помощи во время войны	ЗТ71
	Названия воинских подразделений	ЗТ72
	Понятия, связанные с войной	ЗТ73
7. Время	Временные понятия	ЗТ74
	Дни недели	ЗТ75
8. Религия	Понятия, связанные с обучением и трудом	ЗТ76
9. Учеба и труд	Организация учебы, труда	ЗТ77
	Понятия, связанные с обучением и трудом	ЗТ78
10. Меры	Меры длины, веса, скорости и т.п.	ЗТ79
11. География	Имена населенных пунктов	ЗТ80
	Имена стран, материка, частей света, островов, гор и др.	ЗТ81
	Имена рек, морей, озер, проливов	ЗТ82

Знаковое значение глагольных лексем представляет собой сложный феномен, детерминируемый тремя факторами: 1) соотнесенностью с реальной действительностью, которая представляет собой в данном случае не мир предметов, а мир их отношений, действий и состояний; 2) категориальной семантикой сочетающихся с глаголом предметных имен; 3)

типом смысловых отношений между действием, его субъектом и объектом, а соответственно типом моделей «объект – действие», «действие – объект», «субъект – действие – объект» [6,113].

По сравнению с именами существительными класс глагольных лексем представляет большие трудности при выделении типов по характеру их знакового значения.

Во-первых, собственно лексическое (индивидуальное) значение глагольных слов осложнено целым рядом семантических признаков, выражающих способ протекания действия, обстоятельства его свершения, видовременные характеристики глагольного действия, модальные и экспрессивные наслоения.

Во-вторых, определение глагольной семантики затруднено уточнением, какой из признаков присущ только данной лексеме, какой – целой группировке, а какой обобщен до уровня категории глагольных лексем и т.п.

Традиционно глаголы по их семантике делятся на три класса: 1) глаголы действия: 2) глаголы состояния; 3) глаголы процессов.

У каждой лексико-семантической группы – свои парадигматические параметры, семантические признаки. Так, для глаголов «движения», довольно хорошо изученных во многих языках, выделяются следующие семантические (лексические) категории: направленные движения, скорость движения, манера передвижения, средство или среда передвижения и.т.п.

Для семантики эмотивных глаголов существует иная шкала – глаголы воздействия, глаголы восприятия, глаголы оценки или субъективного отношения и т.п. [6,148].

М. Оразов в своей книге «Қазақ тілінің семантикасы» пишет, что неконкретность лексического значения глагола в основном зависит от его денотативного и сигнификативного значения. Когда речь идет о денотативном значении глагола, мы говорим о движении предмета, физическом действии (амал-эрекет) и о состоянии (қалпы). Если учесть, что движение происходит в какое-то определенное время и в определенном месте, то имеется в виду присутствие субъекта, который осуществляет физическое действие, известна также цель передвижения. При этом лексическое значение глагола дополняют множество других элементов. В качестве примера он приводит следующее предложение: *Шаңғышылар селоға кіргенде көз байланып та қалған еді. (М.Сундетов). /Когда лыжники вошли в село, уже наступило сумерки/.* В этом предложении рассматривается структура значения глагола *кір (входи):* 1) наличие движения; 2) движется некоторый предмет или человек (субъект); 3) движение происходит на земле; 4) переход через некоторую границу (вошли на территорию села); 5) имеется элемент значения войти «внутрь». Перечисленные значения определяются благодаря отношениям и связам слов в данном предложении. Таким образом, действие *кіру (входить)* происходит на лыжне (на снегу) движением лыжников, *кіру* означает вхождение на территорию, где проживают люди (в село). Благодаря сочетанию слова *кір* с другими словами *(село,*

шаңғышылар), мы узнаем направление движения, что оно происходит «извне в село».

Глагол *кір,* сочетаясь с другими словами, может принять абсолютно другое значение. Например, *Мен 1962 жылы партия ұйымына кірдім*. В этом предложении слово *кір* уже не означает движения, а дает понять, что он стал членом партийной организации. Последний пример позволяет говорить о том, что значение слова непосредственно не зависит от значения словосочетаний (от контекста, в широком смысле) в предложении. В первом примере глагол *кір* сочетается с объектом, означающим местность, поэтому его граница является основой понимания лексического значения из сем движения. Во втором примере глагол *кір* сочетается с понятием, т.е. названием организации, здесь сема глагола *кір* единственная – стать членом данной организации. Изменение значения некоторых глаголов требует очень тонкого подхода при анализе их семантики. Например, глаголы движения в одних сочетаниях входят в структуру глагола общения, а в других – в структуру глагола мышления, Сравните: *үйге келу* (прийти домой), *қарсы келу* (возразить), *ойға келу* (вспомнить), *тілге келу* (заговорить) и др. В структуре предложения глаголы образуют отношения с определенными словами. Такие отношения происходят между субъектом или объектом и движением, между физическим действием *(амал-әрекет)* и объектом или субъектом. Когда в предложении глагол требует связи с другим словам, то он с этим словом имеет связь в отношении и значения, и формы (валентность глагола).

Тематические и лексико-семантические классы казахских глаголов. Основанием для отнесения слов в различные тематические и лексико-семантические классы являются обобщающие (жалпылаушы) семы. Необходимо отметить, что в индо-европейских языках имеется определенный опыт в разделении на лексико-семантические классы отдельных частей речи, но в отношении тюркских языков, в частности казахского, подобные исследования начали появляться только в последнее 10-летие (большинство из них имеют сопоставительный с русским языком характер). В настоящее время описаны следующие лексико-семантические группы глаголов казахского языка: глаголов физиологического воздействия [8], глаголов речи [9], глаголов ориентированного положения в пространстве [10], глаголов движения [11], эмотивные глаголы [12], глаголы физического восприятия [13], глаголы созидания [14] и др.

В своей семантической классификации знаменательных классов слов казахского языка старались придерживаться результатов исследований М.Оразова, а именно основных условий объединения слов в одну лексико-семантическую группу, а также и других работ вышеупомянутых ученых.

В центре лингвистической проблематики находится изучение фундаментальных семантических категорий со сложной формально-содержательной структурой. К числу таких категории, отражающих одну из форм всеобщей связи явлений и фактов объективной действительности, следует отнести категорию каузальности (причинности).

В исследованиях казахского языка категория каузальности не была объектом специального изучения. Каузальные отношения в языке рассматривались в пределах отдельно взятых лингвистических задач. Например, в работах А.К.Калыбаевой-Хасеновой, С.В.Шкурдина и др. на морфемном уровне рассматривались глаголы в форме понудительного залога, образующие морфологическую каузативность. На синтаксическом уровне изучались сложноподчиненные и бессоюзные сложные предложения с причинно-следственным характером отношений между составляющими их частями (М.Б.Балакаев, Т.Кордабаев, Н.Х.Демесинова, К.Есенов, Т.Ж.Тажибаева, У.Салиева и др.).

В последнее 10-летие научная школа, возглавлямая доктором филологических наук профессором З.К.Ахметжановой, намного продвинула вперед функционально-семантическое направление исследования в сопоставительном плане изучения казахского и русского языков. Необходимо отметить, что при семантической классификации каузативных глаголов казахского языка мы в своей работе в основном придерживались положений работы Г.М.Алиной «Каузальные отношения в системе русского и казахского языков» (рукопись канд. дис., руков. З.К.Ахметжанова).

Необходимо отметить, что бытийные глаголы казахского языка еще не освоены до конца. В существующих классификациях (И.Маманова, А.Ыскакова, Н.Оралбаевой, М.Оразова, Н.Душимовой и др.) лексико-семантических групп казахского языка бытийные глаголы не выделяются как отдельная группа. Глагол «быть» является основным средством выражения понятия бытия в русском языке. В казахском языке, глаголы с чисто бытийным значением, такие, как *еді (екен, емес), бол,* характеризуются в основном как вспомогательные, которые самостоятельно не употребляются, но образуют описательные формы времен и наклонений основных глаголов.

А. Ыскаков, в свою очередь, определяет глаголы, образованные от корня *-е (еді, екен, емес),* как неполные слова, которые утратили свое лексическое значение либо полностью, либо наполовину и играют в предложении вспомогательную роль, придавая грамматическое значение отдельным словам *(бала еді, қырық-отыздар шамасында екен, қара-торы ала емес).*

Учитывая все высказывания исследователей о семантических классах казахских глаголов, мы для наглядности решили свести их в одну таблицу (***табл. 2***). По нашей классификации семантический класс казахских глаголов делится на 10 разновидностей: I. Глаголы существования; II. Глаголы физических действий; III. Глаголы движения; IV. Глаголы, означающие отношения между людьми, животными и неживыми объектами; V. Глаголы психической деятельности; VI. Глаголы физического состояния и нормы; VII. Глаголы психического состояния; VIII. Фазисные глаголы; IX. Модальныеглаголы; X. Каузативные глаголы.

Каждый класс глагола может быть разделен еще и на семантические подклассы. В данной статье мы решили показать лишь фрагмент, охватив некоторые семантические классы и подклассы казахских глаголов. В таблице

показана следующая информация: 1. В отдельной строке под римской нумерацией дается семантический класс глаголов; 2. В первом столбце таблицы указаны семантические подклассы каждого класса; 3. Во втором столбце – соответствующие примеры к каждому подклассу казахских глаголов; 4. В третьем столбце даны соответствующие коды каждого класса и подкласса, которые будут использованы при записи формул предложений, абзацев и текста.

Таблица 2. **Семантические классы казахских глаголов, используемых в модели порождения текста (фрагмент)**

Семантический подкласс глагола	**Примеры**	**Код**
1	2	3
Семантический класс: I. Глаголы существования		
1.Бытийные глаголы: *Бар*	*Менде жақсы оқуға мүмкіншілік* ***бар****, бірақ оған түк ынтам* ***жоқ.***	ET101
Жоқ	*Оның жасы әлі кәмелетке толған* ***жоқ.***	ET102
Бол (болу, болмау)	*Сөзіңді айтып* ***бол****.* *Кеше жиналыс* ***болды***	ET103
Е (еді, екен, емес)	*Алматы Қазақстанның астанасы* ***еді.*** *Өзі бір қызық адам* ***екен.*** *Оның аты қара-торы* ***емес.***	ET104
Семантический класс: II. Глаголы физических действий		
1. Глаголы физического изменения: созидания и производства, соединения и деления и т.д.	*Жасау, соғу, тігу, тұрғызу, салу, бөлшектеу, екшеу, ыдырату* и т.д.	ET201
2. Глаголы активных физических действий: разрушения и порчи	*Қирату, бүлдіру, бұзу, сындыру, құлату* и т.д.	ET202
3. Глаголы эмоциональных физических действий	*Аузы ашылу, мәз болу, күлу, ыржиу, шапалақтау, секіру (қуанғаннан)*	ET203
4. Глаголы употребляющиеся при приеме пищи, свойственных для человека и животногого мира	*Ішу, жеу, ұрттау, жұту, асау, дәм тату, сораптау* и т.д.	ET204
5. Глаголы прочих физических действий, свойственных человеку и животному миру	*Жүгіру, секіру, ырғу, құсу, таптау, жылжу, қыймылдау ....*	ET205
6. Глаголы физических действий, описывающих неживую природу	*Күркіреу, сарқырау, арсылдау, сілкіну* и т.д.	ET206
7. Глаголы прочих физических действий, свойственных неживой природе	*Ағу, бату (күн), құлау, қайнау, дірілдеу, ысу, суу, гүрсілдеу* и т.д.	ET207
8. Глаголы, связанные со звуками окружающей среды, глаголы звуковыражения	*Тарсылдау, сатырлау, сарқырау, ызыңдау,* *гүрсілдеу, қоңқылдау* и т.д.	ET208
9 Глаголы физических действий, свойственных только животному миру	*Ұлу, үру, ырылдау, мөңку, мөңіреу, жорту, шабу* и т.д.	ET209
Семантический класс: III. Глаголы движения		
1. Глаголы, означающие движения, происходящие на земле	*Жүру, жүгіру, еңбектеу, домалау,сырғанау* и т.д.	ET301
2. Глаголы движения, происходящие на	*Жүзу, сүңгу, малту, қалқу* и т.д.	ET302

воде (в воде)		
3 Глаголы движения, происходящие в воздухе	*Ұшу, самғау, қалықтау, шүйлігу, зымырау* и т.д.	ET303
4. Глаголы движения, означающие направленное движение	*Өрлеу, ылдилау, келу, кету, жақындау, ұзау*	ET304
5. Глаголы движения, означающие хаотичность движения	*Аралау, шарлау, кезу, қыдыру, атқу* и т.д.	ET305
Семантический класс: IV. Глаголы, означающие отношения между людьми, животными и неживыми		
1. Глаголы выравнивания	*Алу, қосу, теңестіру, салыстыру, бөлу* и т.д.	ET401
2. Глаголы принадлежности	*Иемдену, меншіктеу, иелену, баурау и т.д.*	ET402
3. Глаголы приветствия и прощания	*Амандасу, сәлемдесу, көрісу, қоштасу и т.д.*	ET403
4. Глаголы почтения	*Құрметтеу, қадірлеу, ізеттеу, сыйлау и т.д.*	ET404
5. Глаголы властвования	*Басқару, әмір ету, боп-салау, өкім ету и т.д.*	ET405
6. Глаголы подчинения	*Көну, бас иіу, ырқына көну и т.д.*	ET406
Семантический класс: V. Глаголы психической деятельности		
1. Глаголы ощущения: а) глаголы ощущения в общем смысле	*Сезу, сезіну, байқау, абайлау, аңғару, пайымдау, түйсіну ...*	ET501
б) глаголы видения	*Көру, көз салу, көз тастау, үңілу, бажырайу* и т.д.	ET502
в) глаголы вкусовые	*Кермектену, ашқылтымдану, тату*	ET503
г) глаголы осязания (связанные с запахом)	*Бықсу, мүңку, аңқу, сасу, борсу, қаңсу* и т.д.	ET504
д) глаголы осязания, связанные с порами тела человека и животных	*Сыздау, ашу, қышу, тызылдау, сырқырау* и т.д.	ET505
е) глаголы, связанные со слухом	*Тыңдау, есіту, құлақ салу, дауыстау, естілу* и т.д.	ET506
2. Глаголы мышления: а) глаголы мышления, связанные с процессом мышления вообще	*Ойлау, пайымдау, ой жүгірту, ақылға салу, болжау, шамалау* и т.д.	ET507
б) глаголы предположения (приблизительности)	*Шамалау, болжау, тұспалдау, межелеу* и т.д.	ET508
в) глаголы, связанные с выводами, заключениями, обобщениями	*Қорытындылау, шешілу, кесіп айту* и т.д.	ET509
г) глаголы доверия	*Сену, достасу, дос көру, жақындасу, сенім арту.*	ET510
д) глаголы запоминания	*Ескер, ұмытпа, есте сақта, зерделеу қалдыру ...*	ET511
3. Глаголы психического воздействия	*Қуанту, күлдіру, қорқыту, ойландыру* и т.д.	ET512
4. Глаголы внимания и желания	*(жақсы, ақылды, сыпайы, ұстамды)болу* и т.д.	ET513
5 Глаголы активной психической деятельности	*Айқайлау, ұрысу, зар жылау, қарқылдау* и т.д.	ET514

6. Глаголы эмоциональные: а) глаголы восхищения	*Таңдану, қуану, таңырқау и т.д.*	ET515
б) глаголы радости	*Шаттану, қуану, мәз болу ...*	ET516
в) глаголы удовлетворения	*Рахаттану, қанағаттану, шаттану и т.д.*	ET517
г) глаголы эмоционального переживания (скорби)	*Қапалану, қайғыру, қабырғасы қайысу и т.д.*	ET518
Семантический класс: VI. Глаголы физического состояния и нормы		
1. Специфичных для элементов неживой природы (изменение цвета)	*Жылтырау, жарқырау, ағару, қарайу, сарғайу, сәулелену* и т.д.	ET601
2. Глаголы физического сохранения	*Сақталу, бұзылмау, ашымау...*	ET602
а) означающие внутренние психологические состояния человека	*Қасарысу, ашулану, қайғыру, толқу, қуану, ойлану, армандау ...*	ET603
б) означающие субъективные черты человека (наружные)	*Кісімсу, адамсу, баласыну, сасқалақтау* и т.д.	ET604
3. Глаголы, передающие изменение объема предмета, вещества	*Шөгу, ұзару, солу, толысу, зораю, кеңею* и т.д	ET605
4. Глаголы, передающие изменение предмета в количественном отношении	*Азаю, көбею, молаю, ортаю, қосылу* и т.д.	ET606
5. Глаголы, означающие изменение внутренних свойств человека или предмета	*Іру, шіру, үгілу, тозу, босау, өзгеру* и т.д.	ET607
6. Глаголы, употребляющиеся в связи с привычками людей	*Әдеттену, үйрену, дағдылану, жалқаулану ...*	ET608
7. Глаголы, связанные с процессом рождения и роста, свойственных человеку или животным	*Туу, өсу, өну, ержету, жігіт болу* и т.д.	ET609
Семантический класс: VII. Глаголы психического состояния		
1. Глаголы, свойственные только человеку	*Қайғыру, күңірену, ренжу, күлу, жылау, қуану*	ET701
2. Глаголы, связанные с природными явлениями	*Азынау, гуілдеу, борандату, желдету и т.д.*	ET702
3. Глаголы, свойственные человеку и особям животного мира	*Қайғыру, қуану, ашулану, ызалану и т.д.*	ET703
4. Глаголы, приписываемые элементам неживой природы	*Қайғыру, мұңаю, күңірену, қуану, жадырау...*	ET704
5. Глаголы, свойственные человеку, особям животного мира и риписываемые элементам неживой природы	*Қайғыру, мұңаю, жадырау, күңірену, қуану, күлімсіреу и т.д.*	ET705
Семантический класс: VIII. Фазисные и модальные глаголы		
1. Начало действия	*Бастау, кірісу, жасағалы-, кіріскелі-*	ET801
2. Продолжения действия	*Жалғастыру, созу, атқару, -ып, -ip (отыр)...*	ET802
3. Конец действия	*Аяқтау, қою, бітіру, тоқтату ...*	ET803
и.т.д	и.т.д	и.т.д

Далее мы решили опустить подробный табличный вид каузативых глаголов. В нашей таблице они заимствованы из рукописи кандидатской диссертации Г.М.Алиной «Каузальные отношения в системе русского и казахского языков» и делятся на следующие классы и подклассы:

I. Собственно-каузативные глаголы: 1) понуждение; 2) вызывание, причинение; 3) побуждение; 4) обусловливание; 5) воздействие; 6) допущение.

II. Частно-каузативные глаголы: 1) созидания объектов; 2) уничтожения объекта; 3) со значением рождения; 4) со значением лишения жизни кого-либо; 5) со значением изменения объектов; 6) со значением каузации контакта предметов; 7) со значением каузации функтивных отношений; 8) со значением каузации движения; 9) со значением каузации посесивных отношений;

Литература

1. *Оразов М.* Қазақ тілінің семантикасы. Алматы. 1991. -213 б.

2. *Оразов М.* Семантика казахского глагола: дис. докт. филол. наук Алма-Ата. 1983.

3. *Ахметжанова З. К.* Функционально-семантические поля русского и казахского языков. Алма-Ата. 1989. 106 с.

4. *Васильев Л. М.* Семантика русского глагола. М., 1981. - 184 с.

5. *Зубов А. В.* Вероятностно-алгоритмическая модель порождения текста. Автореф дис. док. филол наук. Минск. 1985. 399с.

6. *Уфимцева А.* Типы словесных энаков. М., 1974

7. *Апресян Ю. Д.* Экспериментальное исследование семантики русского глагола М., 1967. 252 с.

8. *Қасабекова Қ. Қ.* Қимыл-қозғалыс етістіктердің жасалу ерекшеліктері //Қазақстан мектебі \\ 1985. № 4. 27-6

9. *Курамысова К. Н.* Семантико-грамматические особенности глаголов речи //Актуальные проблемы общественых наук. Атма-Ата. 1986. Ч. 2. 206 с.

10. *Дүйсекова К. К.* Семантика глаголов ориентированного положения в пространстве в казахском языке. Автореф дисс. Канд. филол. наук. Алматы. 1996 26 с.

11. *Шайкенова Ж.* А. Семантика глаголов движения в казахском и русском языках: Автореф дисс. канд. филол наук. Алматы. 1981. 23 с.

12. *Ибадильдина З. К.* Эмотивная лексика казахского и русского языков: Автореф. дис канд филол. наук. Алматы, 1997 26 с.

13 *Уразымбетова А Б.* Глаголы физического восприятия русского и казахского языков и некоторые аспекты их перевода: Автореф. дис. канд. филол. наук. Алматы. 1998. С. 25.

14 *Токтарова Т. Ж.* Семантика глаголов созидания казахского и русского языков: Автореф. дис. канд. филол. наук. Алматы. 1998. 25 с.

ОСНОВНЫЕ ПРИНЦИПЫ ФОРМАЛИЗАЦИИ СОДЕРЖАНИЯ КАЗАХСКОГО ТЕКСТА

РК МОН Актюбинский государственный
университет им. К.Жубанова. Материалы
Республиканской научно-теоретической

конференции «Жұбанов тағылымы».
Актобе, 4-5 октября 2002 г. С. 40-42.

С провозглашением казахского языка как государственного возрастает его роль в управлении народным хозяйством, обучении и воспитании, культурной жизни и информационном обеспечении суверенной республики.

Являясь высокоразвитой республикой, Казахстан не может остаться в стороне от нового этапа научно-технической революции, связанного с широким внедрением во все аспекты жизни республики компьютерной техники. Если в промышленности такая техника используется достаточно успешно, то информационные технологии на базе казахского языка применяются еще очень слабо.

Эффективное развитие всех областей науки в республике невозможно без создания автоматизированных систем управления народным хозяйством. Основой таких систем в новых условиях должен стать казахский язык. Широкому распространению казахской художественной и научной литературы будет в значительной степени способствовать создание автоматизированных поисковых систем в библиотеках, на предприятиях, в вузах, школах. Естественно, что такие системы должны функционировать на базе казахского языка. Значительно может ускорить процесс активного внедрения в практику казахского языка создание автоматизированных систем обучения в школах и вузах. Как новое средство обучения компьютер вызывает большой интерес, дает значительный эффект при обучении.

Одной из основных причин недостаточного использования компьютеров в информационных технологиях на базе казахского языка и в его обучении является отсутствие формальных описаний казахского языка, которые могли бы быть восприняты электронной машиной. Существующие грамматики казахского языка создавались в свое время для их использования человеком. Правила таких грамматик не формализуемы или плохо формализуемы, так как в основном обращены к интуиции носителя языка, его лингвистическим и экстралингвистическим знаниям. Такими свойствами современные компьютеры не обладают, поэтому необходимо создавать описания языков, в том числе и казахского, которые были бы им понятны.

Создание подобных описаний наталкивается на ряд особенностей, специфичных для казахского языка. К таким особенностям относятся сложность, неоднокодность, недетерминированность и нечеткость языка [1, 3-4]. Сложность казахского языка заключается в наличии в нем большого числа разнородных единиц (морфем, слов и т.д.), способных вступать в огромное число разнообразных отношений. Суть неоднозначности языка в том, что при любом способе квалификации этих единиц число таких классов достаточно велико и разброс количества единиц, входяших в эти классы, также достаточно широк. Недетерминированность казахского языка здесь выражается в том, что отношения между формой единицы и ее содержанием (словом и его значением, предложением и его смыслом, текстом и его

содержанием) многозначны, Наконец, недетерминированиость казахского языка заключается в нечеткости областей значений слов и выражений, в нечеткости границ синтаксической правильности выражений этого языка. Эти особенности казахского языка и создают значительные трудности при построении формальных описаний для его использования в системах автоматического анализа и синтеза казахского текста, а также в автоматизированных системах обучения языку и гуманитарным дисциплинам на базе казахского языка.

В плане традиционного метода исследования грамматического строя казахского языка ученые не ставили перед собой цель применения его в автоматизированных системах, поэтому непосредственное использование в нашей работе имеющихся фактов не всегда представляется возможным. Все же некоторые данные, полученные казахстанкими учеными, брались нами за основу.

Казахстанские ученые при изучении грамматического строя казахского языка сталкиваются с различными проблемами. Следует отметить, что подобное изучение казахского языка, хотя в той или иной мере затрагивалось в первых учебниках и отдельных публикациях, только с 30-40-х годов оно стало приобретать научно-исследовательский характер. В эти годы вышла в свет работа Н.Т.Сауранбаева [2], посвященная сложным предложениям. Некоторые вопросы сложного предложения, а именно классификация придаточных предложений, были проанализированы С.Жиенбаевым [3].

В синтаксической системе любого языка особое место занимает структура простого предложения, ибо общие закономерности, присущие синтаксису, в основном вытекают из природы простого предложения. Эта область синтаксиса была подробно изучена М.Балакаевым. Благодаря его исследованиям, в казахском языкознании появился совершенно новый труд, посвященный синтаксису словосочетаний казахского языка. В этом и в других своих исследованиях ученый сумел доказать, что типы словосочетаний представляют собой особую отрасль синтаксиса языка /[4, 25].

Исследователи казахского языка впоследствии стали выдвигать много интересных проблем, связанных со словосочетаниями и предложениями. В последние годы различные виды предложений исследовались К.Есеновым, согласно данным которого, в предложении конструируются такие главные закономерности, присущие синтаксису, как сочетаемость слов, синтагматические группы, синтаксические группы, синтаксические обороты, связь с другими предложениями и т.д. [5,103].

Ученые в области синтаксиса тюркских языков, изучая актуальные проблемы сложных предложений, стали предлагать включить в парадигму сложных предложений также текст, в составе которого обычно рассматривается и сложное синтаксическое целое, и абзац, и текст с прямой и косвенной речью [6].

Текст – сложная единица. содержащая основные единицы языка: морфемы, слова, словосочетания, предложения. В связи с этим, невозможно

построить модель, позволяющую компьютеру понять содержание текста без разработки правил формализации значений этих основных языковых единиц.

Цель нашей работы – поиск основных принципов формализации содержания казахского текста, как общепринятой единицы общения людей.

Задача нашего исследования состояла в выделении тех семантически значимых единиц казахского текста и правил, которое можно варьировать по определенным вероятностным законам. Полагаем, что рассмотрение в работе основные принципы формализации содержания казахского текста и другие материалы исследования помогут решению некоторых задач по автоматическому анализу и синтезу казахских текстов в информационных, исследовательских и учебных целях.

Литература:

1. Кенесбаев С.К., Бектаев К.Б., Пиотровский Р.Г. О статистическом изучении тюркских текстов. - Статистическое и информационное изучение тюркских языков. Алма-Ата, 1969. - с. 3-4.

2. Сауранбаев Н.Т. Қазақ тіліндегі құрмалас сөйлемдер жүйесі. Алматы, 1948.

3. Жиенбаев С. Синтаксис мәселелері. - Алматы, 1941.

4. Балакаев М.Б. Основные вопросы синтаксиса простого предложения в современном казахском литературном языке. А.Д.Д., Алма-Ата, 1950.

5. Есенов К. Осложненные предложения в казахском языке. - Алма-Ата: Ғылым, 1967.

6. Закиев М.З. Актуальные проблемы в сложных предложениях в тюркских языках. - Баку, 1983.

ФОРМАЛЬНОЕ ОПИСАНИЕ СЛОВАРНОЙ СТАТЬИ АВТОМАТИЧЕСКОГО СЛОВАРЯ КАЗАХСКОГО ЯЗЫКА (АСКЯ)

Институт языкознания им. А.Байтурсынова
МОН РК. «Тілтаным» («Языкознание»).
Алматы-2002, №2. 20-31 стр.

Известно, что словарь модели лингвистической информации о слове строится по машинным основам изменяющихся слов. Такие основы для русского языка, например, выделяются путем отсечения от словоформ окончаний и суффиксов в соответствии с разработанными для данной модели типами формообразований различных классов слов [1]. Для казахского же языка, который по своей типологии и морфологической структуре относится к группе агглютинативных языков, нет особой необходимости выделять такие машинные основы путем формального отсечения, т.к. сам строй данного языка благоприятствует решению этого вопроса. Разработка этой проблемы помогла бы поставить на повестку дня исследовательской

практики более важную проблему – создание теории формальной грамматики казахского языка, которая дала бы ощутимые результаты в компьютерной обработке современных письменных текстов казахского языка. Известно, что имеющиеся академические грамматики казахского языка преследуют во многих случаях иную цель, поэтому создание формальной грамматики казахского языка, понятной компьютерным языкам, должно быть, на наш взгляд, одной из неотложных задач ученых-казаховедов. При этом, как нам кажется, узловой единицей, стержнем вокруг которого будет группироваться языковая информация, должно стать слово, как наиболее естественная единица языка. Такая постановка вопроса обосновывается по нашему, в силу того, что слово можно структурировать и в глубину – на морфемы, фонемы, слога, буквы, значения, а также во вне – как член словоизменительных, словообразовательных, синтаксических, грамматических, ассоциативных и концептуальных рядов [2]. Кроме того, в каждом слове амальгамирована его история, которая связана с историей народа и его культуры. Поэтому-то «слово» — является объектом изучения и лексикологии, и семасиологии, и морфологии, и синтаксиса словосочетании, а также объектом стилистики и др. областей языкознания.

Вышесказанное свидетельствует о том, что формальное описание словарной статьи «Автоматического словаря казахского языка» должно быть организовано так, чтобы всю соответствующую информацию, сконцентрированную вокруг данного слова, в любой момент можно было бы извлечь, а амальгаму – соответствующим образом расшифровать. В основе способа хранения такой информации должно лежать своеобразное «досье» на каждое слово, представленное как подробно структурированная область базы данных, содержащая какое-то число потенциальных сегментов, которые отражают всю систему фонетических, морфологических, синтаксических, семантических, семасиологических, стилистических, историко-этимологических, диалектологических, контекстуальных и т.п. параметров, т.е. всех возможных сведений о слове.

Такие автоматические словари как **АСКЯ** и формальная грамматика, могут быть жизнеспособными и эффективными, если они опираются на общую филологическую традицию и культуру, на глубокое изучение казахского языка и учет информации о нем во всех формах его существования.

Создание модели лингвистической и энциклопедической информации о казахском слове позволит поставить на практические рельсы конкретное моделирование лексико-семантической, словообразовательной, формообразовательной и других сруктур казахского языка.

Мы в своей работе разработали лишь обобщенную схему такой области, содержащую некоторую лингвистическую информацию о казахском слове. Составлением же принципиальной схемы автоматического формирования «досье» и ее реализацией на современных компьютерах, должны заняться специалисты по созданию автоматизированной базы лингвистических данных.

О структуре автоматического словаря казахского языка (АСКЯ). Учитывая, что «слово» любого языка по своей природе является сложным лексико-грамматическим единством, необходимо проводить разделение их на определенные структурные группы или типы, согласно выбранному для данного конкретного случая принципу классификации.

Наш принцип создания **автоматического словаря казахского языка (АСКЯ)** преследует частный случай и не претендует быть универсальным словарем, охватывающим все случаи исследовательской практики.

В самом общем виде можно сказать, что в основе нашего **АСКЯ** лежит алфавитный словарь слов-основ казахского языка, где к каждому слову представлена лингвистическая информация – словарная статья, которая представляет из себя отдельные зоны (области памяти) в которых записаны условно принятые нами кодовая информация о некоторых морфологических, фонетических, семантических, синтаксических и др. признаках (параметрах) рассматриваемого слова. Слово-основ, в данном случае, надо понимать как звуковой комплекс с определенным значением и представляющее собой корневые и производные (аффиксальные) слова, относящиеся к определенному классу слов казахского языка. Слово-основа и есть основная лингвистическая единица **АСКЯ**, к которой приписана соответствующая словарная статья, состоящая из лингвистической (и энциклопедической) информаций, записываемые в виде формально принятых нами кодах в указанном номере зон. Количество зон для каждой единицы словаря будет зависеть от максимального количества «досье» к данному слову. Причем, каждая зона с одним и тем же номером, должна иметь информацию одного и того же типа (содержания). Например, зона №1 – сообщает информацию о принадлежности рассматриваемого слова к определенному классу слов: существительное, глагол, прилагательное и т.д. При этом, некоторые зоны для отдельных слов «временно» могут иметь «нулевую» (пустую) информацию до тех пор, пока его содержание окончательно не выяснится. Это даст возможность постоянно редактировать и расширять объем «досье» к слову.

Схематический **АСКЯ** имеет следующий вид:

Слово-основа	Зона №1	Зона №2	Зона №3	...	Зона№ i	...	Зона№ n

← ..... Словарная статья ..... →

В случае, когда содержание информации для данной зоны – «компактная» в смысле занимаемого объема памяти, то ее можно поместить в соответствующем выделенном номере зоны. Если же такая информации о слове требует дополнительный объем памяти (чем выделено) или же она еще не существует в готовом виде, то такая зона должна иметь соответствующую команду отсылающую на выполнение задания по поиску или созданию такой информации. Например, при синтезировании всевозможных словоформ для данного слова-основы или же при снабжении данного слова контекстными примерами из разных иллюстрационно-текстовых источников и др.

Данный факт говорит о том, что «досье» к слову в автоматическом словаре может быть невсегда доступным потребителю непосредственно, т.е.

невсегда известно в готовом виде. В таких случаях должен быть указан путь достижения цели (алгоритом). Иначе говоря, в этом случае необходимо провести соответствующее исследование по заранее указанному алгоритму и формировать требуемую информацию для данной зоны.

Таким образом, по нашему замыслу, **АСКЯ** должен как бы обладать «пониманием» известных языковых законов, т.е. по возможности в **АСКЯ** должны быть смоделированы синхронные и диахронные языковые процессы, которые представлены в динамике, функционировать отношения единиц, уровней, подсистем и т.д. Следовательно, автоматический словарь казахского языка – **АСКЯ** должен существовать в компьютере не как готовая книга-словарь, а как научная лаборатория, банк научных знаний об объекте (в данном случае о казахском слове). В перспективе, **АСКЯ** должен стать не только как хранилище знаний о языке, а должен представлять из себя как «разумный», специализированный банк знаний, способный предложить решение или нацелить исследователя на определенное решение.

Зона №1. Слово из автоматического словаря принадлежит к одному из ниже перечисленных классов слов.

№ п/п	**К л а с с с л о в а**	**Сокращ.**	**Код**
1	Имя существительное – зат есім	ЗТ	013Т
2	Глагол – етістік	ЕТ	01ЕТ
3	Имя прилагательное – сын есім	СН	01СН
4	Местоимения – есімдік	ЕС	01ЕС
5	Имя числительное – сан есім	СА	01СА
6	Наречия – үстеу	ҮС	01ҮС
7	Служебные слова - шылау	ШЛ	01ШЛ
8	Подражат-ые слова – еліктеуіш	ЕЛ	01ЕЛ
9	Междометие – одағай	ОД	01 ОД

По нашей версии, порядок расположения информации, т.е. информация в соответствующих номерах зон, может быть произвольной, но стандартным для всех единиц словаря и содержание словарной статьи к слову-основе в начальных номерах зон должно выглядеть как показано в таблице – ***Зона №1.***

Зона №2. Слово из автоматического словаря относится одному или нескольким рассматриваемым подъязыкам:

№п/п	**Подъязыки**	**Сокр.**	**Код**
1	Художественный	ХД	02ХД
2	Публицистический	ПБ	02ПБ
3	Научно-популярный	НП	02НП
4	Художественным и публицистический	ХП	02ХП
5	Художественный и научно-популярный	ХН	02ХН
6	Публицистический и научно-популярый	ПН	02ПН
7	Художественный, публицистический и научно-популярный	Х-	02Х-

Зона №3. Слово из автоматического словаря относится к одному из ниже перечисленных:

№ п/н	Способ образования	Сокраще-ния	Код
1	Корневое	КС	03КС
2	Аффиксальное	АС	03АС
3	Слитное	СС	03СС
4	Парное	ПС	03ПС
5	Повторное	ПП	03ПП
6	Сокращенное	СК	03СК

Зона №4. Слоговой состав слова из автоматического словаря состоит:

№ п/п	Слоговой состав слова	Сокраще-ния	Код
	Односложное – бір буынды	ББ	04ББ
2	Двусложное – екі буынды	ЕБ	04ЕБ
	Многосложное – үш, не көп буынды (три и более)	КБ	04КБ

Зона №5. Слово из автоматического словаря относится к соответствующему семантическому классу.

Значение семантических кодов данного слова определяется по принадлежности его к определенному классу слов и из соответствующих классификационных таблиц [13], разработанных нами, которые в данной статье не приводятся.

Таким образом, все сказанное выше по первым пяти зонам на примере парного слова «***Әке-шеше***» в **АСКЯ** должно иметь следующий вид:

Әке-шеше	013Т	02ХД	03ПС	0404	0541

слово-основа ... зоны с № 1 и по № 5 .....

Ввиду того, что порядок расположения информации в словарной статье может быть произвольной, то далее, по усмотрению создателя **АСКЯ**, можно продолжить все предполагаемые лингвистические и энциклопедические информации в их кодовых значениях помещать в оставшихся зонах словарной статьи. Раскодирование информации и поиск необходимой информации в словарной базе, как мы уже говорили, а также их экранный вид – дело специалистов по созданию **АСКЯ**.

В **АСКЯ** обязательно должны быть даны все аффиксальные словоизменения, т.е. всевозможные словоформы от рассматриваемого слова-основы. С этой целью можно было бы составить алфавитный словарь словоформ по разнымм текстам и выбрать по ним все словоформы, соответствующие данной слове-основе. Данная процедура, во-первых, не гарантирует охват всевозможных вариантов формообразования и, во-вторых, постоянное их присутствие в **АСКЯ** заняло бы достаточно большой объем памяти в словарной базе. Исходя из этого мы предлагаем следующий подход.

Для основных классов слов казахского языка (существительное, глагол, прилагательное) воспользоваться грамматическими правилами

формирования словоформ и разработать алгоритм их порождения, т.е. провести синтез казахских словоформ по ее графическим формальным признакам.

Агглютинативный строй казахского языка намного облегчает эту задачу формирования словоформ по известной основе слова, т.к. любое словоизменение образуется посредством последовательного присоединения к основе слова соответствующих аффиксов, а также путем присоединения в виде цепочек одних аффиксов к другим аффиксам.

Допустим, что рассматриваемое в **АСКЯ** слово-основа относится к определенному классу слов и в ***Зоне №6*** необходимо формировать всевозможные словоформы, т.е. перечень словоизменений (грамматические значения) с одним лексическим значением.

Рассмотрим случай, когда слово-основа относится к классу существительных. Как известно, существительное в казахском языке как часть речи объединяет категорию знаменательных слов-имен, которые обладают общими семантическими и грамматическими признаками [4]. Кроме того, данная часть речи является одним из многочисленных по количеству и сложных по структуре. Именно за счет существтельных пополняется словарный состав любого языка, в том числе и казахского.

Морфологическая структура аффиксов существительных и их возможные сочетания были исследованы в количественном отношении и ранее [5]. По статистическим данным в разных стилях казахского языка в среднем около 40% всех слов и 45% всех словоформ относятся только к существительным. Это, видимо, объясняется не только потому, что им присущи различные наименования предметов и явлений, но и от той его функции в предложении, в которых они выступают в роли подлежащего, определения, дополнения, обстоятельства и именного сказуемого и т.д. Имени существительному также присущи характерные только для него аффиксы словообразовании и словоизменения. Мы в данной статье будем говорить только о словоизменительных аффиксах существительных, порождающие словоформы с одним лексическим значением.

Известно, что слово-существительное, вступая в семантико-грамматическую связь с другими словами предложения изменяются по числам, принимая или не принимая аффиксы множественности, по падежам, принимая аффиксы простого или притяжательного склонения и наконец, по лицам, принимая местоименные аффиксы. Последние, т.е. местоименные аффиксы (или сказуемости) присоединяются к существительным только со значениями специальности человека или родственных отношений между людьми. Этот факт учитывается и в нашем исследовании.

А.Ахабаев и К.Б.Бектаев исследовали теоретически возможные формы-варианты присоединения аффиксов к основе существительных, а также частоту их встречаемости в газетных и художественных (роман «Абай жолы») текстах казахского языка. Для удобства изложения укажем принятые им такие условные обозначения: Р – основа слова (корневая и производная); К – аффиксы множественности; С – аффиксы простого склонения; Т –

аффиксы притяжательности; Ж – местоименные аффиксы (или аффиксы сказуемости).

Согласно данным исследования, обозначенные четыре разновидностей аффиксов, присоединяясь к основе слова и в рознь, и сочетаясь в определенной последовательности, теоретически могут встречаться в 496 форм-вариантах. Здесь, конечно, учитывается количественное разнообразие внутри каждой из 4-х разновидностей (К, С, Т, Ж). Из возможных 496 форм-вариантов в газетных текстах были обнаружены только – 225 видов (45%), в художественном тексте – 276 видов (56%). В двух стилях слова-основа Р, т.е. основа с нулевым аффиксом, в среднем составляет 40,54% всех существительных, а оставшиеся 59,46% относятся к существительным со словоизменительными аффиксами.

Какие же последовательности аффиксов присущи существительным казахского языка?

Действительно, и наши исследования, и исследования выше названных авторов подтверждают следующий возможный порядок присоединения аффиксов к основе-слова:

I. Р-С, Р-К-С, Р-К-Т-С, Р-Т-С, Р-Т-К-С;

II. Р-Т, Р-К-Т;

III. Р-К, Р-Т-К, Р-Е-Ж-К;

IV. Р-Ж, Р-Т-Ж, Р-Т-С-Ж, Р-С-Ж, Р-К-Т-Ж, Р-К-Ж.

Как видно из этой записи, аффиксы *К, Т, С, Ж* могут присоединяться к основе *Р* как непосредственно (в рознь), так и сочетаясь в определенном порядке. В данном случае мы специально сгруппировали их в четыре подгруппы так чтобы, в каждом случае замыкал цепочку один из видов словоизменительных аффиксов. Такой порядок расположения подгрупп соответствует занимаемому месту по частотности встречаемости этих подгрупп среди всех существительных со словоизменительными аффиксами. Например, на первом месте стоит подгруппа, когда замыкающим цепочку аффиксов является простое склонение – «С», которые в суммарном подсчете форм составляет более 70% (в указанных стилях – 73% и 78%) всех словоформ. Далее, на II месте – «Т» (21% и 17%), на III месте – «К» (7% и 4%) и на последнем месте, когда замыкающим является – «Ж» (0.01% и 0.3% соответственно в стилях). Таким образом, при словоизменительной процедуре для существительных казахского языка более свойственным в конце словоформы является аффиксы простого склонения.

Наше исследование преследут несколько иную цель, а именно для каждой единицы АСКЯ, т.е. для корневых или производных слов-основ, принадлежащих к классу существительных, формальным путем воссоздать всевозможные аффиксальные словоизменения – словоформы. С этой целью мы за основу своего исследования взяли известные в грамматике правила, которые гласят, что для единственного числа слов в зависимости от последнего слога и звука основы слова простое склонение имен имеет пять фонетических вариантов. Эти пять фонетические варианты нам помогли составить пять типов списков возможных вариантов аффиксов,

используемых для порождения словоформ от корневых и производных основ существительных с конечным твердым слогом и с конечным мягким слогом (без аффиксов сказуемости).

В зависимости от мягкости или твердости последного слога и звука основы слова в АСКЯ аффиксальное словоизменение может принять те или другие падежные окончания. Здесь мы опускаем возможные варианты присоединения к форме единственного числа слов особых аффиксов. Хотим только напомнить, что эти варианты зависят от последнего слога и звука слова в именительном падеже и осуществляются согласно закону сингармонизма.

По морфологическим правилам известно, что слова-существительные могут иметь различные формы и путем притяжательного склонения, т.е. когда указывается принадлежность предмета кому(чему)-нибудь. Необходимо при этом учесть и то, что форма принадлежности указывает одновременно и на предмет обладания и на имя обладателя.

Из-за ограниченности объема статьи, полный список разработанных нами аффиксов, предназначенных для твердых конечных слогов в данной статье не приводятся, но ими можно ознакомиться в журнале «Тілтаным» («Языкознание»). Алматы-2002, №2. 20-31 стр., а также по книге А.К.Жубанова «Основные принципы формализации содержания казахского текста» (Алматы-2002, 195-221-стр).

На основе табличных данных нами разработана и составлена укрупненная блок-схема алгоритма порождения словоформ от корневых и производных существительных с мягким и твердым конечным слогом (см. «Тілтаным» («Языкознание»). Алматы-2002, №2. 20-31 стр.). Благодаря данной блок-схеме можно составить компьютерную программу по синтезу словоформ существительных казахского языка. Нельзя утверждать, что в названных таблицах и в алгоритме программы учтены все случаи аффиксального словоизменения существительных казахского языка. Не исключено, что во время реализации компьютерной программы будут дополнены или исправлены списки для мягких и твердых вариантов аффиксаций существительных.

Подобное исследование проведено нами когда в АСКЯ рассматриваемое слово относится к классу – глагол. Также разработаны и составлены списки аффиксов, используемых для порождения словоформ от корневых и производных основ с конечным твердым и мягким слогом. Нами выявлены 700 вариантов формоизменительных вариантов аффиксации как для глагольных основ с мягким конечным слогом, так и для случая с твердым слогом. Благодаря этим данным мы разработали алгоритм компьютерной программы по реализации синтеза глагольной словоформы.

ЛИТЕРАТУРА

1. Зубов А.В, Переработка текста естественного языка в системе «человек-машина». Л.: Наука, 1971. С. 286-434.

2. Караулов Ю.Н. Методология лингвистического исследования и машинный фонд русского языка // Машинный фонд русскою языка: идеи и суждения, М.: Наука, 1986, С. 13-25.

3. К проблеме семантической классификации знаменательных классов слов казахского языка // Известия МОН РК, НАН РК, N2, март-апрель. Алматы, 2001. С. 13-30.

4. Современный казахский язык. Алма-Ата, Изд. АН КазССР, 1962. С, 130.

5. Ахабаев Ә., Бектаев Қ.Б. Зат есім сөздердің морфологиялық құрылымына статистикалық талдау // Қазақ текстінің статистикасы, Алматы: Ғылым, 1973. 49-87-бб.

УДК 811.512.122: 004.912 *На правах рукописи*

Жубанов Аскар Кудайбергенулы

ОСНОВНЫЕ ПРИНЦИПЫ ФОРМАЛИЗАЦИИ СОДЕРЖАНИЯ КАЗАХСКОГО ТЕКСТА

10. 02. 02 – казахский язык

А В Т О Р Е Ф Е Р А Т
диссертации на соискание ученой степени
доктора филологических наук

Республика Казахстан
Алматы
2002

Работа выполнена в Институте языкознания им. А.Байтурсынова Министерства образования и науки Республики Казахстан

Научный консультант: – доктор филологических наук, профессор **Зубов А.В.**

Официальные оппоненты – доктор филологических наук, профессор **Оразов М.**

– доктор филологических наук, профессор **Жунисбек А.**

– доктор филологических наук, **Момынова Б.**

Ведущая организация – Алматинский государственный университет им.Абая

Защита состоится «29» ноября 2002 г. в 14 часов на заседании диссертационного совета Д.53.38.01 по защите диссертаций на соискание ученой степени доктора филологических наук по специальностям 10.02.06 - тюркские языки и 10.02.02 – казахский язык при Институте языкознания им. А.Байтурсынова МОН РК (480100, г. Алматы, ул. Курмангазы, 29).

С диссертацией можно ознакомиться в центральной научной библиотеке Министерства образования и науки Республики Казахстан (480100, г. Алматы, ул. Шевченко, 28).

Автореферат разослан «____» _____2002г.

Ученый секретарь
диссертационного совета,
доктор филологических наук,
профессор **Ж.А.Манкеева**

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы. Изучение систем высокой степени сложности как живые организмы, общество и язык требует разработки новых теорий, учитывающих многоуровневость их структурной организации. Сказанное, в определенной степени, справедливо и для современной лингвистической науки, так как новые задачи, которые поставила жизнь перед лингвистикой, уже не могут быть решены традиционными методами. К ним относятся существующие способы переработки информации, например, перевод на иностранные языки или реферирование, аннотирование казахских текстов, поиск в них необходимой потребителю информации и т.п. Возрастает роль языка, как основного средства фиксации, хранения и обмена информации различного типа, и, таким образом, на передний план в языкознании выходит коммуникативная функция языка. Центр внимания перемещается к проблеме реализации в языковых процессах законов мыслеобразования, т.е. к языковому процессу порождения текста, как основной единицы письменной коммуникации.

Если в промышленности компьютерная техника используется достаточно успешно, то в обучении и воспитании, культурной жизни и информационном обеспечении суверенной Республики на базе казахского языка такая техника все еще применяется очень редко. Не менее значима еще одна проблема, связанная с процессом обучения, для решения которой необходимы знания правил организации содержания казахского текста. Речь идет об умении выделять в текстах «главное» и «второстепенное», которое необходимо учащимся для правильного и быстрого постижения их содержания, а также может стать основой для составления принципиально новых учебников и учебных пособий.

Одной из основных причин недостаточного использования компьютеров в информационных технологиях и в процессе обучения на базе казахского языка является отсутствие его формальных описаний, которые могли бы быть восприняты электронной машиной. Правила существующих грамматик не формализуемы или плохо формализуемы, так как они, в основном, обращены к интуиции носителя языка, его лингвистическим и экстралингвистическим знаниям.

Создание подобных описаний наталкивается на ряд особенностей, специфичных для многих языков, в том числе и для казахского языка. К таким особенностям относятся сложность, неоднородность, недетерминированность и нечеткость языка.

Эти особенности казахского языка и создают значительные трудности при построении формальных описаний для его использования в системах автоматического анализа и синтеза казахского текста, а также в автоматизированных системах обучения языку и гуманитарным дисциплинам на базе казахского языка.

Ученые в области синтаксиса тюркских языков, изучавшие актуальные проблемы сложных предложений, считают возможным включить в

парадигму сложных предложений также текст, в составе которого обычно рассматривается и сложное синтаксическое целое, и абзац, и текст с прямой и косвенной речью. Известно, что текст – сложная единица, включающая основные единицы языка: морфемы, слова, словосочетания, предложения. В связи с этим, построить модель, позволяющую компьютеру понять содержание текста без разработки правил формализации значений этих основных языковых единиц, невозможно.

Цели и задачи исследования. Цель диссертационной работы – разработка основных принципов формализации содержания казахского текста, как общепринятой единицы общения людей, семантически значимых единиц казахского текста и правил, понятных компьютеру и объединяющих эти единицы в связный текст.

В соответствии с целью исследования, поставлены следующие задачи:

1. Изучить лексико-семантическую организацию казахских текстов, рассматривая объектом исследования следующие три подъязыка, представляющих три наиболее употребительных функционально-смысловых типов речи – описание, рассуждение и повествование: художественные тексты (рассказы известных казахских писателей); тексты научно-популярного стиля; публицистические (газетные) тексты.

2. Изучить логико-семантическую организацию абзацев указанных текстов.

3. Исследовать роль вероятностных и детерминированных правил в организации содержания абзаца и текста.

4. Разработать специальный язык, понятный компьютеру, позволяющий формально представлять содержание предложения, абзаца и текста.

5. Найти способы формального выделения основного статического содержания казахского текста.

6. Составить принципиальный алгоритм, позволяющий компьютеру создать текст по заданному основному статическому содержанию.

Научная новизна. Научная новизна проведенного исследования заключается в том, что впервые предложена формальная процедура представления содержания казахского текста на двух взаимосвязанных уровнях: статических единиц содержания (опорные слова) и динамических единиц содержания (семантико-синтаксических формул предложений и абзацев). Для записи таких формул предложен специальный семантико-синтаксический язык (СЕМСИНТ).

Впервые проведена классификация абзацев по их функционально-смысловому, структурному и предметно-логическому содержанию.

Разработан принцип семантической классификации лексики казахского языка на уровне знаменательных классов слов (существительное, глагол, прилагательное и т.д.) и условного кодирования семантических подклассов, предназначенный для формальных процедур при записи формул предложений, абзацев и текста в целом.

Впервые предлагается метод формального формообразования с помощью специально составленных списков аффиксов, синтез словоформ от слов-основ существительных и глагольных основ, а также алгоритм указанной задачи. Рассмотрены возможности создания автоматического словаря казахского языка (АСКЯ).

Одним из важных результатов данного исследования является составление принципиального алгоритма перехода от естественного казахского текста к его формальному представлению.

Основные положения, выносимые на защиту:

1. Содержание некоторого текста может быть представлено в виде некоторой неоднозначной последовательности более мелких единиц плана содержания (микротем, подтем, субтем и т.д.).

2. Минимальной лингвистической единицей, несущей такой «квант» содержания, является абзац.

3. Тема текста (его статика) может быть представлена в виде небольшого числа наиболее употребительных существительных, глаголов и прилагательных (главных и второстепенных опорных слов).

4. Развертывание темы в связный текст (динамика текста) осуществляется путем неоднозначного подбора с помощью детерминированных правил и случайных чисел абзацев с определенным предметно-логическим содержанием, записанным на специально разработанном семантико-синтаксическом языке.

Методы исследования. Задачи, поставленные в диссертации, решались различными методами, которые, учитывая сложность решаемой задачи, применялись комплексно:

– метод лингво-статистического анализа, с помощью которого получены частотные списки слов по текстам 3-х стилей и лексико-морфологические характеристики этих текстов;

– логико-семантический и статистический методы для выделения микротем в текстах;

– дистрибутивно-статистический метод для изучения организации содержания текста из отдельных микротем.

Помимо указанных, использовались также такие, как метод теории алгоритмов, компонентный, сопоставительный, описательный, контекстный и другие методы.

Теоретическая и практическая ценность работы. Теоретическое значение настоящего исследования заключается в создании системы, правил, позволяющих представить содержание казахского текста в формализованном виде, понятном современному компьютеру. Эти правила могут стать основой алгоритмов и программ для автоматического анализа и синтеза казахского текста.

Практическая значимость исследования заключается в возможности создания на основе разработанных принципов и алгоритмов системы компьютерных программ для автоматического анализа и синтеза казахского текста. Предложенный формальный критерий выделения опорных слов,

может стать основой для определения информативных и ключевых слов при построении компьютерных систем обучения казахскому языку.

Апробация результатов работы. Материалы диссертации были доложены на следующих Всесоюзных и Республиканских научных конференциях и рабочих совещаниях: на первой научной конференции молодых ученых АН КазССР. – Алма-Ата, 1968; на Всесоюзной конференции по статистическому и информационному изучению тюркских языков. –Алма-Ата, 1969; на Всесоюзной научной конференции по структурно-математическим методам моделирования языка. – Киев, 1970; на второй научной конференции молодых ученых АН КазССР, – Алма-Ата, 1970; на Всесоюзной конференции по автоматической переработке текста методами прикладной лингвистики. – Кишинев, 1971; на Всесоюзной школе-семинаре по автоматическому распознаванию слуховых образов. – Алма-Ата, 1973; на третьей научно-теоретической конференции молодых ученых АН КазССР. – Алма-Ата, 1974; на Всесоюзной тюркологической конференции: «Советская тюркология и развитие тюркских языков в СССР». – Алма-Ата, 1976; на научном семинаре по статистической оптимизации преподавания языков и инженерной лингвистике. – Чимкент, 1980; на Всесоюзной конференции по переработке текста методами инженерной лингвистики. – Минск, 1982; на рабочем совещании по машинному фонду языков народов СССР. – Тбилиси, 1988; на рабочем совещании по машинному фонду языков народов СССР. – Таллин, 1988; на третьей Всесоюзной конференции по созданию машинного фонда русского языка. – Москва, 1989; на I Всесоюзной конференции: «Тюркская фонетика-90». – Алма-Ата, 1990; на Всесоюзной конференции по актуальным проблемам компьютерной лингвистики в Тартуском университете. – Тарту, 1990; на Всесоюзной конференции по использованию ЭВМ в научной и учебной работе гуманитарных вузов. – Минск, 1990; на Республиканских Жубановских чтениях, посвященных актуальным проблемам современных гуманитарных наук: «Жұбанов тағылымы». – Актобе, 1998; на Республиканской научно-теоретической конференции «Жұбанов тағылымы». Министерство науки и высшего образования РК, Актюбинский университет им. К.Жубанова. – Актобе, 1999; на Республиканской научно-теоретической конференции, посвященной 100-летию К. Жубанова: «Құдайберген Жұбанов және қазақ тіл білімі». – Алматы, 1999; на Республиканской научно-практической конференции, посвященной 100-летию К.Жубанова: «Мемлекеттік тіл: терминология, іс қағаздары мен бұқаралық ақпарат құралдарының тілі». – Астана, 1999; на Межвузовской научно-практической конференции 3-4 февраля 2000 года: «Актуальные проблемы теории и практики перевода в контексте современности». – Алматы, 2000; на Республиканской научно-теоретической конференции «Жұбанов тағылымы-V». Министерство науки и высшего образования РК, Актюбинский университет им. К.Жубанова. – Актобе, 2002.

Публикации. По материалам диссертации опубликовано 55 работы, в том числе, 2 монографии, 3 частотных словаря (в соавт.), 23 в материалах и тезисах конференций (9 зарубежных).

Структура и объем диссертации. Диссертационная работа состоит из двух книг.

Первая книга содержит основной материал диссертации, изложенный на 250 страницах машинописного текста, иллюстрирована 37 таблицами, 5-ю рисунками и 2-я блок-схемами алгоритмов программ. Она включает введение, пять разделов, заключение, список цитированной литературы из 327 наименований работ отечественных и зарубежных авторов.

Вторая книга в объеме 310 страниц, содержит Приложения к основной части диссертации, включающие конкретные лингво-статистические материалы, использованные в процедуре определения основных принципов формализации содержания казахского текста.

1 Текст и его основные знаковые характеристики

Такие процессы, как управление, накопление и переработка информации на основе широкого использования электронно-вычислительной техники, требуют решения ряда задач, связанных с сознательным конструированием новых знаков и знаковых систем.

В этой связи, в разделе 1 представлен обзор данных литературы, включающий следующие подразделы: 1.1 Язык, речь. Основные функции и единицы; 1.2 Текст как основной объект современной лингвистики; 1.3 Текст и его единицы как знаки. Подходы к определению «знак»; 1.4 Предложение как речевой знак; 1.5 Сверхфразовое единство и абзац как знаки; 1.6 Текст как знак.

В разделе даны точки зрения ученых, начиная с Ф. де Соссюра и завершая мнениями современных ученых, на понятия «язык» и «речь», а также на определенные границы между их функциями.

Ввиду того, что наше исследование проводится на уровне абзацев, остановимся на семиотическом представлении абзаца.

Для построения семиотической схемы речевого знака-абзаца необходимо отметить следующие положения: а) разные части абзаца несут различную информационную нагрузку; б) абзац в целом отражает нечто общее в ряду подобных отношений предметов действительности; в) абзац может вносить в текст определенные стилистико-экспрессивные элементы.

Итак, все перечисленные свойства абзаца дают нам возможность представить знаковую ситуацию абзаца так, как это показано на ***рисунке 1***.

Согласно представленной нами семиотической структуре знака-абзаца, содержанием абзаца будет являться отражение в нашем сознании отношения линейной последовательности имен абзаца к отраженной в сознании человека микроситуации как совокупности семи компонентов означаемого этого знака.

Изучая структуру абзаца по текстам различных типов и сопоставляя цепочки предложений, не принадлежащих к одному абзацу, можно выявить особенности абзаца и включенных в него предложений:

а) предложения, входящие в абзац, должны описывать одну микротему (или несколько взаимосвязанных микротем), т.е. в составе денотативных и коннотативных значений слов, пронизывающих абзац, должны находиться общие компоненты;

Семиотическая структура знака-абзаца

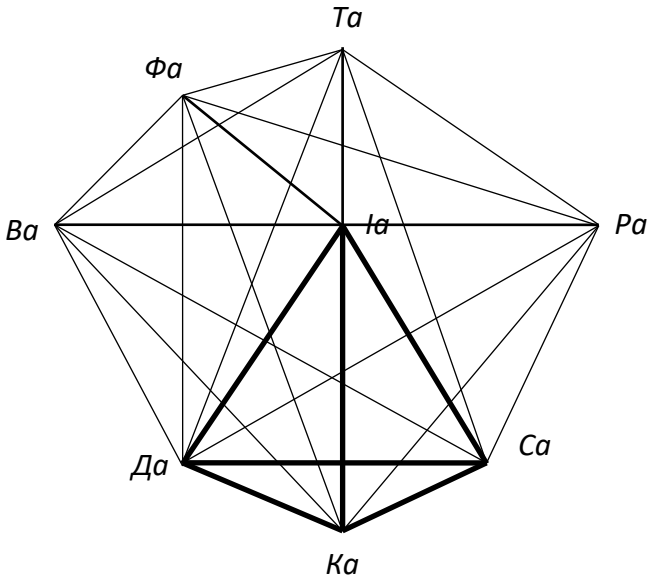


Рисунок 1. *Да* – денотат абзаца; *Са* – десигнат; *Ка* – коннотат; *Іа* – цепочка из «n» имен абзаца; *Ва* – синтаксический аспект абзаца; *Ра* – информационный аспект; *Та* – тематический аспект; *Фа* – связующий межфразовый аспект абзаца.

б) предложения одного абзаца должны иметь специальные средства связи между собой;

в) между предложениями одного абзаца должны существовать определенные синтаксические связи, отражающие социально утвердившуюся логику, в соответствии с которой автор организует восприятие читателя.

Итак, резюмируя точки зрения лингвистов, психолингвистов, психологов и других специалистов на структуру текста и его составляющие, связь текста с языком, мыслью и реальной действительностью можно сделать следующие выводы:

1. Основными характерными чертами текста являются: наличие единого содержания; наличие определенных единиц текста.

2. Текст можно рассматривать как знак динамической системы, в котором выделяются три плана:

а) план выражения;

б) план содержания;

в) план интерпретации текста потребителем.

2 Формальное представление словаря казахского языка

2.1 Индивидуальный словарь носителя языка и его роль в восприятии и понимании текста

Наблюдая за частотой различных речевых явлений, можно судить о наличии или отсутствии статистических закономерностей в системе языка в целом или в каких-то его подсистемах.

В традиционном плане частотные словари позволяют глубже исследовать лексическую организацию языка, а именно выделить лексику отдельных функциональных стилей, их общую часть, разграничить национальную лексику от заимствованной, общеупотребительную от терминологической и т.д.

Выбирая наиболее частые и информационно насыщенные единицы языка, а также типовые контексты их употребления, можно построить такую модель языка (базовый язык), которая будет являться некоторым приближением к реальной системе подъязыка, порождающей тексты узкой тематики [Пиотровский Р.Г., 1970].

Пока еще не издан частотный словарь казахского языка в целом, но составлены частотные словари по отдельным жанровым разновидностям. Отметим только то, что покрываемость казахских текстов имеет различия, например, от индоевропейских языков. Если в последнем 100–150 словоформ покрывают 50% обследуемого текста, то в казахском подъязыке публицистики и в художественном стиле такой процент покрываемости текста соответствует 700–800 словоформам.

Такое различие в процентах покрываемости казахских словоформ К.Б.Бектаев объясняет тем, что «... в тюркских языках агглютинативная конструкция по своим функциям часто идентична не столько словоформе, сколько словосочетанию в индоевропейских языках» [Бектаев К.Б., 1978] и слова имеют более разнообразные виды формообразования по сравнению с языками флективного строя. Действительно, в казахском языке только у существительных количество формообразующих аффиксов достигает 500, а большинство казахских глаголов имеют до 1000 различных форм [Бектаев К.Б., Ахабаев А., Мырзабеков С., 1973].

Таким образом, сравнительно небольшая группа наиболее частых слов в сумме составляет огромное число словоупотреблений. Последнее дает возможность, сделать следующий вывод об объеме индивидуального словаря носителя казахского языка: для понимания 80% казахского текста достаточно хорошо знать значения казахских слов в пределах от 2 тыс. до 3 тыс. слов из высокочастотной зоны частотного словаря, составленного по различным жанрам этого языка.

Так как задача формализации содержания казахского текста будет решаться в целях использования современных компьютеров, то возникает ряд вопросов. В частности, какой минимальный объем словаря казахского языка необходимо записать в память компьютера? По каким областям знаний (подъязыкам) лучше всего выбрать экспериментальные тексты? Какие слова

считать «опорными», «главными», вокруг которых сконцентрировано содержание текста? Эти и другие вопросы, касающиеся раскрытия принципов формальной структуры содержания казахских текстов, будут рассмотрены нами детальнее в следующих разделах и подразделах работы.

2.2 Лексико-морфологическая характеристика компьютерного словаря казахского языка

Для более полного охвата различных областей знания нами были взяты тексты из художественной литературы, публицистики и научно-популярной литературы, т.е. эксперименту подвергнуты тексты, относящиеся к трем функционально-смысловым типам речи, таким, как описание, повествование и рассуждение.

Указанные тексты были подвергнуты компьютерной обработке и на начальном этапе были получены следующие частотные словари: 1) алфавитно-частотный словарь-словоуказатель для каждого типа текстов; 2) алфавитно-частотный словарь лексем в целом по каждому стилю с указанием абсолютной частоты слова и количества текстов данного стиля; 3) алфавитно-частотный словарь слов для отдельно взятого номера текстов (ij) для каждого типа текстов; 4) частотно-алфавитный словарь слов по отдельно рассматриваемым типам текстов; 5) обратный алфавитно-частотный словарь слов для каждого стиля.

Результаты статистического исследования, с точки зрения употребительности в них различных классов слов по рассматриваемым трем стилям подтверждают известные теории о специфичности стилей казахского языка. Так, существительные в публицистике и научно-популярной литературе употребляются гораздо чаще, чем в художественной литературе. Относительно глагольных слов казахского языка можно отметить обратное, т.е. глаголы в художественной литературе употреблены чаще, чем в текстах двух других стилей. Эти факты вполне объяснимы: в художественных текстах идет некоторое повествование, развитие действий, выражаемых глаголами. Для текстов научных и публицистических характерны рассуждение, объяснение, которые чаще выражаются существительными. Однако, необходимость напоминания о небольшом числе главных действующих лиц в художественном тексте приводит к возрастанию в них числа имен собственных и местоимений.

В едином словаре, полученном в результате объединения данных всех трех словарей, можно отметить некоторые особенности, свойственные казахскому языку. В частности, соотношение количества глаголов и существительных в едином словаре составляет 0,62, что ближе к тексту научно-популярному, объединяющему черты художественного и научного текста, а в научно-популярном стиле равно 0,55, тогда как в художественном и публицистическом такое соотношение глаголов и прилагательных – 1,1 и 0,49 (подробнее о статистике частей речи в отдельных стилях см.: А.К.Жубанов «Основные принципы формализации содержания казахского текста», 2002).

2.3 Автоматический словарь казахского языка

2.3.1 Формальное описание словарной статьи

Для выделения, так называемых, машинных основ слов казахского языка и решения других подобных задач приходится заниматься проблемой, касающейся теории формальной грамматики казахского языка. При этом, как мы полагаем, узловой единицей, стержнем, вокруг которого группируется языковая информация, должно стать **слово,** как наиболее естественная единица языка и, в силу того, что его можно структурировать и в глубину – на морфемы, фонемы, слоги, буквы, значения, и во вне – как члена словоизменительных, словообразовательных, синтаксических, грамматических, ассоциативных и концептуальных рядов [Караулов Ю.Н., 1986]. Кроме того, в каждом слове амальгамирована его история, которая связана с историей народа и его культуры.

Все это свидетельствует о том, что формальное описание словарной статьи, в рассматриваемом нами «Автоматическом словаре казахского языка» (АСКЯ), должно быть организовано так, чтобы всю соответствующую информацию, сконцентрированную вокруг данного слова, в любой момент можно было бы извлечь, а амальгаму – соответствующим образом расшифровать. В основе способа хранения такой информации должно лежать своеобразное «досье» на каждое слово, представленное как подробно структурированная область базы данных, содержащая какое-то число потенциальных сегментов, которые отражают всю систему фонетических, морфологических, синтаксических, семантических, стилистических, историко-этимологических, диалектологических, контекстуальных и т.п. параметров, т.е. всех возможных сведений о слове.

Такие автоматические словари, как АСКЯ и формальная грамматика могут быть жизнеспособными и эффективными, если они опираются на общую филологическую традицию, культуру, на глубокое изучение казахского языка и учет информации о нем во всех формах его существования. Создание модели лингвистической и энциклопедической информации о казахском слове позволит практически осуществить конкретное моделирование лексико-семантической, словообразовательной, формообразовательной и других структур казахского языка.

Нами разработана обобщенная схема такой области, которая содержит некоторую лингвистическую информацию о казахском слове. Составлением же принципиальной схемы автоматического формирования «досье» и ее реализацией на современных компьютерах должны заняться специалисты по созданию автоматизированной базы лингвистических данных.

2.3.2 Структура автоматического словаря казахского языка

Наш принцип создания автоматического словаря казахского языка (АСКЯ) преследует частный случай и не претендует быть универсальным словарем, охватывающим все случаи исследовательской практики.

В основе нашего АСКЯ лежит алфавитный словарь слов-основ казахского языка, где к каждому слову представлена лингвистическая информация – словарная статья, которая представляет собой отдельные зоны (области памяти), в которых записана кодовая информация о некоторых морфологических, фонетических, семантических, синтаксических и других признаках рассматриваемого слова. Слово-основу, в данном случае, следует понимать как звуковой комплекс с определенным значением, представляющий собой корневые и производные слова, относящиеся к определенному классу слов казахского языка. Слово-основа и есть основная лингвистическая единица АСКЯ, к которой приписана соответствующая словарная статья, состоящая из лингвистической (и энциклопедической) информации, записываемой в виде формально принятых нами кодов в указанном номере зон.

Количество зон для каждой единицы словаря будет зависеть от максимального количества «досье» к данному слову. Причем, каждая зона с одним и тем же номером должна иметь информацию одного и того же типа (содержания). Например, зона 1 сообщает информацию о принадлежности рассматриваемого слова к определенному классу слов: существительное, глагол, прилагательное и т.д. При этом некоторые зоны для отдельных слов «временно» могут иметь «нулевую» (пустую) информацию до тех пор, пока его содержание окончательно не выяснится. Это даст возможность постоянно редактировать и расширять объем «досье» к слову.

Если какая-то информация о слове еще не существует, то такая зона должна отослать пользователя (по соответствующей команде) на выполнение задания по поиску или созданию такой информации. Иначе говоря, в этом случае необходимо провести соответствующее исследование по заранее указанному алгоритму и формировать требуемую информацию для данной зоны. Таким образом, АСКЯ должен как бы обладать «пониманием» известстстных языковых законов, т.е. по возможности в АСКЯ должны быть смоделированы синхронные и диахронные языковые процессы, представленные в динамике. Следовательно, автоматический словарь казахского языка должен существовать в компьютере не как готовая книга-словарь, а как научная лаборатория, банк научных знаний об объекте (о слове), т.е. он должен стать не только хранилищем знаний о языке, но и представлять собой «разумный», специализированный банк знаний, способный предложить решение или нацелить исследователя на определенное решение.

Содержание словарной статьи к слову-основе в начальных номерах зон по нашей версии должно выглядеть следующим образом:

Зона 1. *Слово в АСКЯ имеет условный код, означающий принадлежность его к определенному классу слов (части речи).*

Зона 2. *Слово в АСКЯ имеет условный код, означающий принадлежность его к одному или нескольким рассматриваемым подъязыкам (художественный, публицистический, научный* и т.д.*).*

Зона 3. *Слово в АСКЯ имеет условный код, означающий способ его образования (корневое, аффиксальное, парное, слитное* и т.д.).

Зона 4. *Слово в АСКЯ имеет условный код, означающий его слоговой состав (односложный, двусложный* и т.д.*).*

Зона 5. *Слово в АСКЯ имеет (согласно классификационной таблице) условный код, означающий его семантический класс (подкласс).*

Таким образом, все изложенное по первым пяти зонам на примере парного слова *«әке-шеше»* в АСКЯ будет иметь следующий вид:

әке-шеше	01ЗТ	02ХД	03ЕЕ	0404	053Т059

слово-основа ..... зоны с 1 по 5 .....

Так как, порядок расположения информации в словарной статье может быть произвольным, то далее, по усмотрению создателя АСКЯ, можно продолжить все предполагаемые лингвистические и энциклопедические информации.

2.4 Синтез казахских словоформ-существительных для АСКЯ

В АСКЯ обязательно должны быть даны все аффиксальные словоизменения, т.е. всевозможные словоформы от рассматриваемого слова-основы. С этой целью, используя грамматические правила формирования словоформ для основных классов слов казахского языка (существительное, глагол) мы разработали алгоритм их порождения, т.е. провели синтез казахских словоформ.

К примеру: пусть в ***зоне 6*** необходимо сформировать всевозможные словоформы от слово-основы из АСКЯ, которая относится к определенному классу слов, а именно, к классу существительных.

По статистическим данным в разных стилях казахского языка, в среднем, около 40% всех слов и 45% всех словоформ относятся только к существительным. Это, видимо, объясняется не только тем, что им присущи различные наименования предметов и явлений, но и их функциями в предложении, согласно которым они выступают в роли подлежащего, определения, дополнения, обстоятельства, именного сказуемого и т.д.

Известно, что слово-существительное, вступая в семантико-грамматическую связь с другими словами предложения, изменяется по числам, принимая или не принимая аффиксы множественности, по падежам, принимая аффиксы простого или притяжательного склонения, и наконец, по лицам, принимая местоименные аффиксы.

Как подтверждают наши результаты исследований других авторов, возможный порядок присоединения аффиксов к основе-слову, относящемуся к классу существительных, имеет следующий порядок:

I. Р-С, Р-К-С, Р-К-Т-С, Р-Т-С, Р-Т-К-С;
II. Р-Т, Р-К-Т;
III. Р-К, Р-Т-К, Р-К-Ж-К;
IV. Р-Ж, Р-Т-Ж, Р-Т-С-Ж, Р-С-Ж, Р-К-Т-Ж, Р-К-Ж.

Как видно из этой записи, аффиксы К (көптік жалғау), Т (тәуелдік), С (септік), Ж (жіктік жалғау) могут присоединяться к основе Р (түбір) как непосредственно (в рознь), так и сочетаясь в определенном порядке. Такой порядок расположения подгрупп соответствует занимаемому месту по частоте их встречаемости. Как видно, при этом для существительных казахского языка более свойственны в конце словоформы аффиксы простого склонения.

За основу исследования нами взято известное в грамматике правило, которое гласит, что для единственного числа слов, в зависимости от последнего слога и звука слова-основы, простое склонение имен имеет пять фонетических вариантов.

В целях удобства нами сделана незначительная перегруппировка конечных букв основы, которая не согласуется с общепринятой в единственном случае, а именно, когда основа, оканчивающаяся на звонкие согласные *ж, з,* склоняется в творительно-соединительном падеже.

Для слов-существительных, которые по своему значению связаны с людьми и их деятельностью, профессией и т.д. формоизменяющиеся аффиксы рассматривались нами особо.

Таким образом, нами анализировались следующие четыре случая: 1) основа оканчивается на мягкий слог и среди присоединяемых аффиксов присутствие аффиксов сказуемости необязательно; 2) основа оканчивается на мягкий слог и среди присоединяемых аффиксов участие аффиксов сказуемости обязательно; 3) основа оканчивается на твердый слог и среди присоединяемых аффиксов присутствие аффиксов сказуемости необязательно; 4) основа оканчивается на твердый слог и среди присоединяемых аффиксов участие аффиксов сказуемости обязательно.

Итак, число всевозможных случаев аффиксации именной основы по первому фонетическому варианту в списках оказались равным 130, в других фонетических вариантах – меньше или равно 130, в списках с обязательным участием аффиксов сказуемости – 47.

2.5 Об алгоритме программы – синтез именных словоформ казахского языка

Алгоритм программы синтеза именных словоформ казахского языка составлен с расчетом на четыре указанных выше списков аффиксов и его работа доведена до формальной стыковки выбранной именной основы из АСКЯ с аффиксами из соответствующих упомянутых списков. При этом, учитывали законы сингармонизма и правила прогрессивной и регрессивной ассимиляции, а также некоторые исключения, связанные с перегруппировкой типов основ. Составленная таким образом, укрупненная блок-схема алгоритма задачи синтеза именных словоформ казахского языка послужит основой для составления компьютерных программ для АСКЯ.

2.6 Синтез казахских глагольных словоформ для АСКЯ

Статистическое исследование казахских глагольных словоформ на структурно-морфологическом уровне с использованием материалов частотных словарей по разным стилям казахского языка обращают внимание на статистику в отношении количества и порядка присоединения формообразующих аффиксов к глагольной основе. Результаты таких исследований нами были учтены в процессе синтеза глагольных словоформ.

В модели АСКЯ будут фигурировать глагольные основы, главным формальным признаком которых является возможность прибавления в конце основы отрицательной частицы (***-ма, -ме, -ба, -бе, -па, -пе)***.

Мы предлагаем формальный метод формообразования глаголов, т.е. грамматические категории, такие, как видовые формы, наклонения, время, спряжения глагола, формы причастия и деепричастия (за исключением залоговых форм) синтезировать программным путем, согласно грамматическим правилам современного казахского языка.

Руководствуясь теми положениями, что в казахском языке формообразующие и словоизменяющиеся аффиксы присоединяются к слову в определенном порядке и разработанной нами методикой формообразования существительных, мы составили списки аффиксов, используемых для порождения словоформ от корневых и производных основ казахского глагола.

При этом нами экспериментально выявлено 700 вариантов формоизменительных аффиксов как для глагольных основ с мягким конечным слогом в отдельности, так и для случая с твердым слогом.

3 Семантическая классификация знаменательных классов слов казахского языка

3.1 Принципы семантической классификации лексики казахского языка

При семантической классификации лексики казахского языка, т.е. при выделении в ее составе определенных лексико-семантических классов и подклассов имеются свои трудности и своя специфика. Количество таких классов и подклассов и их состав зависят от целей исследования, лингвистической школы и ряда других факторов. Работ, посвященных исследованию казахского языка в области семантической классификации, относительно мало. Здесь стоит отметить монографию М.Оразова «Қазақ тілінің семантикасы» [Оразов М., 1991].

Контрастивное изучение языков приобретает сегодня не только актуальное теоретическое, но и принципиальное значение для практики обучения второму языку. В этом направлении определенную научную ценность имеют интерлингвальные исследования Э.Д.Сулейменовой по установлению семантических параметров слова в аспекте контрастивного анализа [Сулейменова Э.Д., 1996]. Особого внимания заслуживают также научные изыскания З.К.Ахметжановой, посвященные функционально-

семантическим полям русского и казахского языков [Ахметжанова З.К., 1989]. Эти труды необходимы были нам для решения практических задач нашего исследования.

При семантической классификации лексики казахского языка (в самом общем виде) мы руководствовались следующими тремя принципами: 1) денотативно-десигнативный; 2) парадигматический; 3) синтагматический.

Денотативно-десигнативный принцип учитывает онтологическое расчленение действительности и при этом слова объединялись в группы в соответствии с тем, что они обозначают в реальной действительности («названия животных», «названия растений», наименования определенных классов предметов и т.п.).

При парадигматической классификации слова были объединены в группы на основе тождества определенных семантических компонентов, выделенных в значении слов («модальные глаголы», «каузативные глаголы» и т.п.).

По синтагматическому принципу слова объединяются в лексико-семантические группы с учетом количества и качества валентности классифицируемых слов.

Исходя из целей нашего исследования, в основу семантической классификации положен денотативно-десигнативный принцип, хотя в некоторых случаях он будет дополняться двумя другими, перечисленными выше.

3.2 Семантическая классификация имен существительных казахского языка

Именами существительными в казахском языке, так же, как и во многих тюркских языках, являются различные наименования предметов и понятии в широком смысле слова. Как пишет М.Оразов, среди слов казахских существительных самыми частыми по сложности их значений являются нарицательные имена. Эти имена существительные по отношению к значению имеют в языке изменчивый характер, т.е. иногда у них превалирует денотативное значение, а иногда сигнификативное или приобретает как бы суммарное значение – и денотативное, и сигнификативное. Примером может служить использование слов, которые применяются в метафорическом и метомических значениях.

В количественном отношении тематические группы в любом языке, в том числе, и в казахском, довольно велики. С учетом целей нашего исследования, анализируемого материала, а также отдельных идей, изложенных в научной литературе, нами выделены семантические классы существительных, которые представлены нами в виде классификационной таблицы. В основу семантической классификации казахских существительных нами положено деление реальной действительности на определенные зоны, сегменты. При этом, денотативно-десигнативный принцип использован здесь при выделении наиболее крупных семантических

классов существительных. Более мелкие лексико-семантические группы, например «части тела человека», «воинские звания», «дни недели» и др., связаны парадигматическими отношениями.

Таблица 1

Семантические классы существительных, использованных в модели порождения казахского текста

Сегменты реальн. действительти	Наименование семантического класса	Код класса
1	2	3
Человек и его мир	Человек и его части тела	ЗТ301
	Части лица и головы человека	ЗТ302
	Состояния и ощущения, специфичные для человека	ЗТ303
	Действия, специфичные для человека	ЗТ304
	Понятия, используемые при характер-ке человека	ЗТ305
	Предметы одежды и обуви человека и понятия, с этим связанные	ЗТ306
	Предметы-артефакты, используемые человеком в целях гигиены, для украшения и т.д.	ЗТ307
	Объекты, используемые для передвижения человека и понятия, с этим связанные	ЗТ308
	Предметы питания человека	ЗТ309
	Понятия, связанные с жилищем человека	ЗТ310
	Понятия, связанные со смертью и захоронением	ЗТ311
	Имена, отчества и фамилии людей	ЗТ312
	Национальность человека	ЗТ313
	Предметы-сосуды, которыми пользуется человек, и связанные с этим понятия	ЗТ314
	Предметы, используемые на отдыхе, охоте, рыбалке, и понятия, с этим связанные	ЗТ315
	Наименование гражданских специальностей, должностей, степеней, званий	ЗТ316
	Предметы-артефакты, использование при учении, труде	ЗТ317
	Понятия, связанные с взаимоотношением людей	ЗТ318
	Понятия, связанные с родственными отношениями между людьми	ЗТ319
	Награды человека	ЗТ320

В нашей классификационной таблице реальная действительность подразделена на 11 сегментов: наука; природа; человек и его мир; животный мир; растительный мир; время; религия; война и понятия, с ней связанные; учеба и труд; меры (длины, веса, скорости и т.д.); география. Каждый из перечисленных сегментов в свою очередь подразделены на соответствующие подклассы, суммарное число таких подклассов для 11 сегментов реальной действительности в классификационной таблице равно 75. Например, для сегмента «Наука» число таких классов, включающих наименования, связанные с понятиями различных научных отраслей – 21. Отдельным классам или подклассам в таблице присвоен индивидуальный условный код, который будет использован в модели порождения казахского текста. Пример: «ЗТ101» – код общефилософского понятия; «ЗТ117» – код понятия,

связанного с устройством отдельного механического объекта (*көпір – мост*) и т.д.

В ***таблице 1*** приведен фрагмент из классификационной таблицы существительных, где рассматривается только один сегмент реальной действительности – «Человек и его мир», который внутренне подразделен на 20 семантических классов с кодовыми значениями от «ЗТ301» по «ЗТ320».

3.3 Семантическая классификация глаголов казахского языка

3.3.1 Знаковое значение глагольных лексем

Значения существительных более автономны и абсолютны, а значения глагольных лексем относительны и детерминируются многими факторами. Поэтому соотнесенность глагольных лексем с предметным рядом предполагает наличие смысловых связей глагола с предметными именами. Знаковое значение глагольных лексем представляет собой сложный феномен, детерминируемый соотнесенностью с реальной действительностью, которая представляет собой в данном случае не мир предметов, а мир их отношений, действий и состояний а также типом смысловых отношений между действием, его субъектом и объектом, а соответственно типом моделей «объект – действие», «действие – объект», «субъект – действие – объект» [Уфимцева А.А., 1974].

3.3.2 Тематические и лексико–семантические классы казахских глаголов

Основанием для отнесения слов к различным тематическим и лексико-семантическим классам являются обобщающие семы. Необходимо отметить, что исследования по тематическим и лексико-семантическим классификациям казахских глаголов имеет недавнюю историю и большинство из них имеет сопоставительный с русским языком характер.

В исследованиях казахского языка категория каузальности (причинности) в ее полном объеме не была объектом специального изучения. Каузальные отношения в казахском языке рассматривались в пределах отдельно взятых лингвистических задач[1].

При классификации каузативных глаголов мы пользовались некоторыми фактами работы Г.М.Алиной «Каузальные отношения в системе русского и казахского языков».

---

[1] На морфемном уровне: формы повелительного наклонения глагола (Н.Сауранбаев, А.Исхаков, А.Калыбаева, С.Исаев, Н.Оралбаева и др.; предложные формы глагола (В.А.Исенгалиева, Г.А.Мейрамов, Р.Г.Баяминова и др.); на синтаксическом уровне (М.Б.Балакаев, Т.Кордабаев, Н.Х.Демесинова, К.Есенов и др.).

Традиционно разграничение глаголов по их семантике осуществляется на три следующих класса: 1) глаголы действия; 2) глаголы состояния; 3) глаголы процессов.

Учитывая изложенные выше подходы к классификации глагольной лексики мы предложили следующие 9 семантических классов глаголов, используемых в модели порождения казахского текста: 1) глаголы существования, т.е. бытийные глаголы (коды подклассов: ET101÷ET104); 2) глаголы физических действий (ET201÷ET209); 3) глаголы движения (ET301÷ET305); 4) глаголы, означающие отношения между людьми, животными и неживыми (ET401÷ET406); 5) глаголы психической деятельности (ET501÷ET522); 6) глаголы физического состояния и нормы (ET601÷ET609); 7) глаголы психического состояния (ET701÷ET705); 8) фазисные и модальные глаголы (ET801÷ET807); 9) каузативные глаголы (ET901÷ET944).

По концу сегмента кодового значения семантического класса можно получить следующую информацию, например, по ‘ET944’: первые две буквы ‘ET’ - *етістік* – глагол; первая цифра ‘9’ означает вид семантического класса – каузативные глаголы; последние две цифры ‘44’ – порядковый номер семантического подкласса.

Приведенный фрагмент (***табл. 2***) из классификационной таблицы глагола охватывает только два из перечисленных 9 видов.

3.4 Семантические классы причастий *(есімше)* и деепричастий *(көсемше)* казахского языка

Причастия и деепричастия являются особыми формами глагола и представляют собой своеобразный разряд слов, поэтому при их классификации на лексико-семантические группы были использованы, в основном, те же принципы, что и при классификации глаголов с учетом способа их образования.

В семантическом плане казахские причастия рассматривались нами с точки зрения как физических, так и психических, а также физическо-психических и эмоциональных **действий и состояний.**

1) деепричастия, обозначающие физические действия; 2) ... физические состояния; 3) ... психические действия; 4) ... психические состояния, свойственные человеку.

Семантические подклассы деепричастий выделялись на основании тех же принципов как и в случае с причастием, но с учетом их способа образования и соответствующими кодовыми значениями (в обоз. «КШ» – сокращ. от *көсемше*).

Таблица 2

Семантические подклассы казахского глагола, означающие физические действия, которые используются в модели порождения казахского текста

Семантический подкласс глагола	**Примеры**	**Код**
Семантический класс: 2. Глаголы физических действий		
1. Глаголы физического изменения: созидания и производства, соединения и деления и т.д.	*Жасау, соғу, тігу, тұрғызу, салу, бөлшектеу, екшеу, ыдырату* и т.д.	ET201
2. Глаголы активных физических действий: разрушения и порчи	*Қирату, бүлдіру, бұзу, сындыру, құлату* и т.д.	ET202
3. Глаголы эмоциональных физических действий	*Аузы ашылу, мәз болу, күлу, ыржиу, шапалақтау, секіру*	ET203
4. Глаголы, употребляющиеся при приеме пищи, свойственные для человека и животных	*өшу, жеу, ұрттау, жұту, асау, дәм тату, сораптау* и т.д.	ET204
5. Глаголы прочих физических действий, свойственные человеку и животным	*Жүгіру, секіру, ырғу, құсу, таптау, жылжу, қыймылдау* и т.д.	ET205
6. Глаголы физических действий, описывающих неживую природу	*Күркіреу, сарқырау, тарсылдау, сілкіну* и т.д.	ET206
7. Глаголы прочих физических действий, свойственных неживой природе	*Ағу, бату (күн), құлау, қайнау, ысу, суу, гүрсілдеу* и т.д.	ET207
8. Глаголы, связан. со звуками окружающей среды, гл-ы звуковыражения	*Тарсылдау, сатырлау, сарқырау, ызыңдау, гүрсілдеу* и т.д.	ET208
9. Глаголы физических действий, свойственных только животным	*Ұлу, ұру, ырылдау, мөңку, мөңіреу, жорту, шабу*	ET209
Семантический класс: 4. Глаголы, означающие отношения между людьми, животными и неживыми		
1. Глаголы выравнивания	*Алу, қосу, теңестіру, салыстыру, бөлу* и т.д.	ET401
2. Глаголы принадлежности	*Иемдену, меншіктеу, иелену, баурау* и т.д.	ET402
3. Глаголы приветствия и прощания	*Амандасу, сәлемдесу, көрісу, қоштасу* и т.д.	ET403
4. Глаголы почтения	*Құрметтеу, қадірлеу, ізеттеу, сыйлау* и т.д.	ET404
5. Глаголы властвования	*Басқару, әмір ету, бопсалау, өкім ету* и т.д.	ET405
6. Глаголы подчинения	*Көну, бас иіу, ырқына көну…*	ET406

Таблица 3. Семантические классы причастий *(есімше),* спользованных в модели порождения казахского текста

Семантический подкласс причастий *(есімше)*	**Способ образования и коды**		
	Прошед. время на *-ған /-ген,-қан/-кен*	**Будущее время на *-ар/-ер/-р***	**Будущее время на *-атын/-етін, -йтын/-йтін***
Сем. кл.: 1. Причастия, означающие признаки физических действий			
1. Свойственные чел-ку, особям живот-го мира и элем. нежив. природы	ЕШ001	ЕШ041	ЕШ081
2. Физическо-психических и эмоциональ-ных действий, свойств-х только человеку	ЕШ002	ЕШ042	ЕШ082
3. Специф. для элем-тов неживой природы	ЕШ003	ЕШ043	ЕШ083
4. Свойственные только человеку	ЕШ004	ЕШ044	ЕШ084
5. Свойственные человеку и особям животного мира	ЕШ005	ЕШ045	ЕШ085
Сем. кл.: 2. Причастия, означающие признаки физических состояний			
1. Свойственные человеку, особям живот-ного мира и элементам неживой природы	ЕШ011	ЕШ051	ЕШ091
2. Свойственные элементам только неживой природы	ЕШ012	ЕШ052	ЕШ092
3. Свойственные человеку и элементам неживой природы	ЕШ013	ЕШ053	ЕШ093
4. Свойственные только человеку	ЕШ014	ЕШ054	ЕШ094
Сем. кл.: 3. Причастия, означающие признаки психических действий			
1. Признаки психических и эмоционально-психических действий, свойственных только человеку	ЕШ021	ЕШ061	ЕШ101
2. Признаки психических действия-мышления, свойственных человеку	ЕШ022	ЕШ062	ЕШ102
3. Признаки психических действий, припис. элементам нежив. природы	ЕШ023	ЕШ063	ЕШ103
4. Свойственные человеку, особям животного мира и приписываемые элементам неживой природы	ЕШ024	ЕШ064	ЕШ104
Сем.кл.: 4. Причастия, означающие признаки психических состояний			
1. Признаки психических и эмоционально-психических состояний, специфичных для человека	ЕШ031	ЕШ071	ЕШ111
2. Специфичные человеку, особям животного мира и приписываемые элементам неживой природы	ЕШ032	ЕШ072	ЕШ11
3. Признаки психических состояний, припис. элементам нежив. природы	ЕШ033	ЕШ073	ЕШ11
4. Обоз-ие признаки псих-кого состояния-мышления, свойственные человеку	ЕШ034	ЕШ074	ЕШ11
5. Признаки психических состояний, свойственные человеку и особям животного мира	ЕШ035	ЕШ075	ЕШ11

3.5 Семантические классы прилагательных казахского языка

Специфические формы и семантические особенности имени прилагательного в казахском языке четко отграничивают его от других частей речи. Одна из особенностей казахского прилагательного – наличие морфологических форм, т.е. степени сравнения, а также его препозитивность. Исследователи различных языков почти единогласно делят прилагательные на качественные и относительные. Однако, выделение внутри этих классов более мелких лексико-семантических групп проводится неоднозначно с использованием самых различных принципов и критериев.

Отмечается, что качественные прилагательные крайне неоднородны, так как они, в основном, освобождены от предметного значения (абстрагированы). Как правило, их дифференцируют по их роли в высказывании, в то время как относительные прилагательные могут иметь любые значения, связанные с предметами и событиями. Все это вынуждает нас искать такую семантическую классификацию прилагательных, которая была бы удобна для цели порождения казахского текста и в то же время была относительно небольшой по числу семантических классов.

Предлагаемая нами семантическая классификация прилагательных опирается на денотативно-десигнативный принцип. Естественно, что она не претендует на полноту, но вполне достаточна для решения поставленной перед нами задачи.

Качественные прилагательные казахского языка в семантическом плане нами были подразделены на следующие четыре класса (сегмента), формулировка которых начинается с выражения – «**Качественные прилагательные, обозначающие признаки**» и далее: 1) свойственные человеку и тому, что связано с ним; 2) свойственные особям животного мира; 3) используемые для описания предметов; 4) качественные прилагательные, обозначающие оценочные признаки, свойственные человеку, особям животного мира и используемые для описания предмета.

Семантические классы внутренне подразделены еще на семантические подклассы (от 4 до 10) и они рассматриваются в трех видах, как положительная степень *(жай шырай),* сравнительная степень *(салыстырмалы шырай)* и превосходная степень *(күшейтпелі шырай).* По этой причине количество кодовых значений для отдельного подкласса увеличивается в три раза. Буквенные символы «СС» – сокращение от *сапалық сын есім.*

В ***табл. 4*** представлен фрагмент классификационной таблицы качественных прилагательных, в котором рассмотрен только один семантический класс с его соответствующими подклассами.

Относительные прилагательные *(қатыстық сын есімдер)* казахского языка мы классифицировали на семь семантических классов, рассматривая их **как прилагательные, характеризующие объект по его отношению:** 1) к какому-либо одушевленному субъекту; 2) к какому-либо неодушевленному

предмету; 3) ко времени; 4) к материалу; 5) к какому-либо действию; 6) к месту; 7) к отвлеченному понятию.

В классификационной таблице семантические классы относительных прилагательных рассматривались в сравнительной и превосходной степенях *(салыстырмалы және күшейтпелі шырайлар),* не подразделяясь на подклассы.

Таблица 4

Качественные прилагательные, обозначающие признаки, используемые для описания предметов (семантический класс 3)

Семантический подкласс качественных прилагат-ных	Виды, коды и примеры		
	Положитель. степень	Сравнитель. степень	Превосход. степень
1. Означающие цвета предметов	СС021 *ақ, көк, қызыл*	СС061 *ақшыл ...*	СС111 *аппақ, сүп-сүр*
2. Означающие вкусовые качества пищевых продуктов	СС022 *тәтті, ащы, тұщы* и т.д.	СС062 *тәттілеу, ащырақ. . .*	СС112 *өте тәтті, тым ащы...*
3. Означающие объем предметов	СС023 *дәу, ірі, кіші..*	СС063 *үлкенірек ...*	СС113 *тым ірі ...*
4. Означающие вес предмета	СС024 *ауыр, жеңіл.*	СС064 *ауырлау . . .*	СС114 *өте ауыр . . .*
5. Означающие природное состояние предмета	СС025 *қалың,сирек, селдір, ну ...*	СС065 *қалыңдау, сирегірек*	СС115 *тым сирек, сеп-селдір . . .*
6. Означающие признак предмета, в связи с температурными изменениями окружающей среды	СС026 *ыстық,суық, салIын* и т.д.	СС066 *ыстықтау, суығыраI ...*	СС116 *ып-ыстық, сүп-суық . . .*
7. Употребляющиеся в связи с внешней формой предмета	СС027 *сопақ, үшкіл*	СС067 *сопақтау...*	СС117 *дөп-дөңгелек*
8. Употребляющиеся в связи с природными явлениями	СС028 *жарық, ашық*	СС068 *ашығырақ...*	СС118 *жап-жарық*
9. Употребляющиеся в связи с колич-и отношениями предметов	СС029 *аз, көп, мол*	СС069 *аздау, көптеу*	СС119 *өте аз,шап-шақ*
10. Употребляющиеся в связи со стоимостью предмета	СС030 *арзан, қым-бат, тегін ...*	СС070 *арзанырақ, қымбаттау*	СС120 *ап-арзан, теп-тегін . . .*

3.6 Семантическая классификация наречий

Известно, что казахские слова-наречия по своему значению не однородны, так как одно и то же наречие в зависимости от контекста может иметь разное значение. При семантической классификации принято делить их на **определительные наречия *(анықтауыш үстеулер),*** выражающие качество, количество, меру, образ, манеру действия, способ и совместность действия и на **обстоятельственные наречия *(пысықтауыш үстеулер),*** которые указывают на время, место, причину и цель действия.

Сравнением статистических данных выявлено, что чаще остальных в казахских текстах встречаются наречия образа и способа действия, затем наречия времени, места, количества и меры, реже – наречия с оттенком

усиления, выражающие качество предмета, самого действия, кроме того, меры или объема, как с оттенком усиления, так и ослабления. В группе наречий образа и способа совершения действия наряду с наречиями степени усиления и ослабления действия в отдельную семантическую группу можно отнести и наречия совокупности *(топтау үстеулері)*.

Нами разработана приемлемая для нас семантическая классификация наречий казахского языка, использованная для построения модели порождения казахского текста (см. ***табл. 5***).

Таблица 5

Семантические классы наречий, использованные в модели порождения текста

Семантический подкласс наречия	**Примеры подкласса наречий**	**Код**
1	2	3
Сем. класс: 1. Определительные наречия *(Анықтауыш үстеулер)*		
1. Определительные наречия образа действия (*бейне, сын)*	*Осылай, осылайша, сөйтіп, өйтіп, бүйтіп, олай-бұлай, бірден, бірдей, бірге, бірте-бірте, біртіндеп, ойша, өзінше, іле, қолма-қол….*	YC011
2. Определительные наречия количества и меры *(мөлшер үстеуі)*	*Онша, сонша, осынша, оншалық, соншалық, осыншалық, оншама, мұншама, анағұрлым, біре-сарен, ондаған, көптеген, бірталай ...*	YC012
3. Определит. наречия качества *(сапалық)*: а) с оттен-м усиления или осл-я *(күшейту, әлсірету)*	*Ең, әбден, ылғи, кілең, сәл, өңкей, тіпті, тым, нақ, нағыз, әнтек, мүлдем, дәл, керемет, қабағат, мейлінше, сонша, төтенше, жөнсіз, орасан, өрен, аса, өте ….*	YC013
б) определ. наречия с оттенком собирательности или разделительности *(топтау, саралау)*	*Екеулеп, үшеулеп, он-ондап, бір-бірлеп, он-оннан, қырық-елуден, аз-аздан, көп-көптен, топ-тобымен, рет-ретімен, бас-басына, үйді-үйіне* и т.д.	YC014
Сем. класс: 2. Обстоятельственные наречия *(Пысықтауыш үстеулер)*		
1. Обстоятельственные наречия времени *(мезгіл үстеулері)*	*Бүгін, былтыр, ертең, таңертең, кешке, қазір, енді, әлі, ендігәрі, екіншәрі, бұрын, әуел-баста, ерте, кеш, күндіз, бүрсігүні* и т.д.	YC021
2. Обстоятельственные наречия причины *(себеп-салдар үстеулері)*	*Жоққа, босқа, құр босқа, бекерге, амалсыздан, лажсыздан, шарасыздан, сөйле-сөйлей, көре-көре, оқый-оқый* и т.д.	YC022
3. Обстоятельственные наречия цели *(мақсат)*	*Әдейі, әдейілеп, жорта, қасақана, кітапқа бола, оқуға бола* и т.д.	YC023
4. Обстоятельственные наречия места *(мекен үстеулері)*	*ілгері, ілгеріде, әрі, әріде, әрмен, жоғары, төмен, жолшыбай, жол-жөнекей, алға, артта, алды-артынан, тысқары* и т.д.	YC024

3.7 Семантическая классификация местоимений *(есімдік)*

По семантическим свойствам местоимения в языкознании относят к дейктическим словам. Самыми распространенными в казахском языке являются личные *(жіктеу есімдігі)* и указательные *(сілтеу есімдігі)*

местоимения. В отличие от существительных, прилагательных и числительных, местоимения в языке имеют максимально общее и вместе с тем всеобъемлющее значение и выполняют указательную функцию, обозначая предметы или признаки, они не называют и не характеризуют отдельные конкретные предметы, признаки или определенные количества, а лишь указывают на них.

Известно, что употребление местоимений тесно связано с употреблением имен и является как бы их синонимами. Тем самым, автор текста старается избежать частого повтора одного и того же слова. Таким образом, денотативное и сигнификативное значении местоимения зависят от очень большого числа факторов.

Местоимения являются одним из основных средств связи предложений в абзаце и абзацев в единый текст. Помимо этого, как это будет показано, личные местоимения выступают, как правило, в роли контекстных синонимов главных и второстепенных опорных слов текста. Учитывая отмеченную значимость местоимений в структуре текста и тот факт, что общее число местоимений в казахском языке относительно не велико, появляется возможность полностью охватить их при семантической классификации. В таблице семантических классов мы вводим для каждого местоимения индивидуальный семантический код.

Таким образом, в данном разделе, анализируя конкретные материалы, содержащиеся в 102 текстах трех функциональных типов казахского языка, в плане семантической классификации знаменательных классов слов, нами осуществлены следующие процедуры: 1) оценена статистическая особенность исследованных научных, публицистических и художественных текстов, а также показана их специфика, важная для задач создания единой модели порождения казахского текста; 2) построен единый словарь модели (объединенный алфавитно-частотный, распределительный по типам словарь слов), состоящий из 7240 слов; 3) классифицирована лексика словаря по семантическим сегментам действительности.

4 Абзац как основной семантико-синтаксический компонент казахского текста

4.1 Роль и специфика абзацев в общей структуре текста

Абзац – это реально существующая единица в общей структуре текста, сознательно выделяемая автором для облегчения восприятия и усвоения его мыслей читателем. Содержание всего текста включает содержание входящих в него абзацев. При этом, основным критерием, объединяющим их в единый текст, будет служить тематический критерий. Предложения, входящие в абзац, описывают одну или несколько взаимосвязанных микротем.

В абзацах можно выделить элементарные составляющие типа «действие», «состояние», «место», «время», «средство», «причина», «цель» и т.д. Этим элементарным составляющим микроситуации (ЭСМС) в тексте абзаца будут соответствовать слова или словосочетания (глаголы, причастия,

существительные и их сочетания). Таким образом можно выявить специфику каждой микроситуации, а именно она будет задаваться в виде определенной последовательности ЭСМС и индивидуальной особенностью отдельного абзаца.

Итак, в рассматриваемых текстах можно выявить возможные типы абзацев и каждому абзацу приписать в определенном порядке группы слов-конкретизаторов, замещающих в абзаце соответствующие денотаты микроситуаций.

4.2 Фразовые стереотипы и абзац

Как утверждают ученые, стереотипность, или стандартность речевого действия, играет в основном роль экономии усилия как автора, так и воспринимающего речь индивида. Стереотип усреднен исторически из индивидуальных решений и находится в рамках нормативной грамматики [Kade O. 1980].

Такие «фразовые стереотипы» или синтаксические единицы в современной теории текста называют «сверхфразовыми единствами» или «сложными синтаксическими целыми». Многими исследованиями допускается, что и абзац письменного текста является «логическим слепком» с действительности. Для научных текстов эти правила носят строгий и стереотипный характер, тогда как для художественных текстов – более свободный и творческий.

Объединение в единое целое указанных выше «кусков» информации возможно по следующим трем принципам: 1) вероятностному, если объединение происходит на основе вероятностных распределений; 2) детерминированному, если порядок следования «кусков» однозначен и строго задан; 3) вероятностно-алгоритмическому, если в процессе объединения учитывают и случайные, и детерминированные (алгоритмические) связи.

Таким образом, абзац как смысловой «квант» и как «застывший» синтаксический шаблон, может быть принят за минимальный логико-семантический компонент текста.

4.3 Функционально-смысловая классификация казахского текста

Основная особенность текста любого языка, в том числе, и казахского, заключается в том, что между предложениями (и их комбинациями) устанавливаются смысловые связи на основе отношений между объектами и явлениями действительности. При этом одни предметы или явления действительности объединяет их последовательность во времени, другие – пространственно или иным образом. Ученые отмечают, что в тексте предложения интегрируются в типы высказывания или типы речи (описание, повествование, рассуждение), каждый из которых имеет определенное

функционально-смысловое значение и стабильную структурно-языковую характеристику [Нечаева О.А., 1974].

В случае, когда фразы или более крупные текстовые единицы расположены в тексте в порядке, определяемом связью во времени описываемых в тексте событий, тип связи называют повествованием *(баяндау).* Если же связи между фразами или более крупными единицами текста осуществляются на основе пространственных, функциональных или иных отношений между частями какого-либо отражаемого в речи объекта, тип речи называется описанием *(суреттеу).* Описание – это констатация наличия или отсутствия каких-либо признаков или явлений.

Важным для нас представляются те структурно-смысловые разновидности, которые выделяются в описании: 1) различные описания природы; 2) описание внутренней части какого-либо помещения; 3) описание внешнего вида какого-либо одушевленного существа или неодушевленного предмета; 4) описание-характеристика какого-либо одушевленного существа или неодушевленного предмета.

И, наконец, когда фразы или более крупные единицы текста связаны причинно-следственными выводными отношениями, то такой тип речи называется рассуждением *(пайымдау, ойжүгірту).*

По нашим статистическим данным, наиболее употребительны следующие шесть видов рассуждений: 1) оценочные именные рассуждения; 2) рассуждения со значением состояния; 3) рассуждения с обоснованием реальных или гипотетических действий; 4) рассуждения со значением необходимости, возможности или желательности некоторого действия; 5) рассуждения с заранее обусловленными действиями; 6) рассуждения с категорическим отрицанием или утверждением. Представленные особенности повествования, описания и рассуждения наиболее ярко проявляются в пределах наименьших семантико-синтаксических единиц, т.е. абзацев.

В диалоге каждую смену реплики без косвенной речи автора и косвенную речь, сопровождающую реплику (вместе с репликой), мы решили считать за абзац. Поэтому, в дальнейшем, примем, что существуют абзацы с вопросительной речью, абзацы с побудительной речью и смешанного типа, когда в одном и том же абзаце присутствуют признаки двух или более перечисленных функциональных типов.

Итак, различаются следующие типы абзацев: 1) повествовательные *(баяндау);* 2) описательные *(суреттеу);* 3) типа рассуждения *(пайымдау, ойжүгірту);* 4) абзацы с выражением вопроса *(сұрау мәнді);* 5) абзацы с выражением побуждения *(ой тастау, әсер ету мәнді);* 6) абзацы смешанного типа *(аралас мәнді).*

Известно, что начальные абзацы текста являются некоторым прологом к последующим, излагая точку зрения автора на ту или иную проблему. Конечные или заключительные абзацы, как правило, подводят итог изложенному или служат своеобразным послесловием, позволяющим до конца уяснить мысль автора.

Выявляя особенности функционально-смысловой связи в рассматриваемых казахских текстах, нами получены следующие статистические данные:

1. По художественному типу изучена статистика 20 текстов, состоящих из 487 абзацев. Из 44 начальных абзацев 30 (68%) относятся к типу повествования, 9 (20,4%) абзацев относятся к смешанному типу. Тип повествования свойственен и медиальным, и заключительным абзацам – 54% и 69% соответственно. В художественных текстах более употребительный смешанный вид абзацев и далее, в порядке убывания, следуют абзацы со значением побуждения, вопроса, рассуждения и описания. Для 39 конечных абзацев свойственен тип повествования (69%), на втором месте – побуждения (15,4%), на третьем – рассуждения (10,3%).

2. Статистика 35 газетных текстов показывает, что функционально-смысловой тип – повествование имеет преимущество в начальных и в конечных абзацах – 89% и 96% соответственно. В медиальных абзацах, можно встретить все виды функционально-смысловых типов с характерным перевесом типа повествования.

Из рассмотренных 57 конечных абзацев газетных текстов 55 относятся к типу повествования – это 96% всех конечных абзацев.

3. Рассмотрим статистику функционально-смысловых типов абзацев в 47 научо-популярных казахских текстах. Функционально-смысловой тип – повествование имеет преимущество во всех рассматриваемых структурных типах (т.е. начальных, медиальных и конечных абзацах). Округляя данные, отметим, что 70% начальных, конечных и около 50% всех медиальных абзацев относится к типу повествования. Далее, среди начальных абзацев второе место принадлежит смешанному виду абзацев (17%).

В заключении отметим, что каждый функционально-смысловой тип, в большей или меньшей степени, может встретиться в тексте как в отдельности, так и в смешанном виде, частота их встречаемости в структурном отношении не стабильна, т.к. она зависит от многих других факторов.

4.4 Классификация абзацев казахского текста по их предметно-логическому содержанию

Выделение групп предложений в абзацы не является произвольной операцией автора текста, а определяется рядом факторов: необходимостью изложения новой мысли, важностью передаваемой информации в рамках всего текста, невозможностью дальнейшего линейного представления новых сведений, сменой речевых субъектов и т.д. [Лосева Л.М., 1980]. Рассмотренные в предыдущем пункте функционально-смысловые типы абзацев относятся к общим типовым способам изложения содержания. Дальнейшей конкретизацией содержания абзаца является понятие предметно-логического содержания абзаца [Гришина О.Н., 1982].

Уточняя имеющие в литературе определения «предметно-логического содержания абзаца», примем их в следующей формулировке:

1. Предметно-логическим содержанием абзаца-описания называется дифференцированное по типам главных и второстепенных субъектов, объектов, мест и времен микроситуации, описанной в абзаце, перечисление фактов объективной действительности, связанных с содержанием всего текста.

2. Предметно-логическим содержанием абзаца-рассуждения называется дифференцирование по типам главных и второстепенных субъектов, объектов, мест и времени ситуации, представленной в тексте, последовательная связь мыслей, связанных с содержанием всего текста.

3. Предметно-логическим содержанием абзаца-повествования можно назвать дифференцированное по типам главных и второстепенных субъектов, объектов, мест и времен ситуации, представленной в тексте, последовательное изложение в абзаце действий, событии, состояний с детализацией или без нее.

Руководствуясь изложенными определениями, нами классифицированы и выявлены некоторые статистические особенности абзацев в трех рассматриваемых типах текстов по их предметно-логическому содержанию.

4.5 Особенности абзацев научно-популярного текста по их предметно-логическому содержанию

Исследователи отмечают, что научный текст на всех его уровнях членится с опорой на «логический слепок» с действительности на основе обобщения человеческого опыта научного познания.

На основе научно-популярных текстов, которые нами взяты для эксперимента, лежит определенная «однородность» текста в тематико-ситуационном, пространственном и временном плане.

Рассматривая 93 начальных абзацев в 47 научно-популярных текстах по их предметно-логическому содержанию, нами было выявлено 20 более характерных типов начальных абзацев (см. ***табл. 6***).

Как видно из ***таблицы 6***, наиболее употребительный из типовых начальных абзацев это тот, который по существу отражает наиболее типичное содержание начальной части научно-популярного текста, представляющего главный объект исследования и то, что сделано или будет сделано автором для раскрытия сути этого объекта.

Подобный же анализ 353 медиальных абзацев научно-популярного текста позволил нам выделить 23 типовые разновидности по предметно-логическому содержанию абзацев.

В научно-популярных текстах более употребительными среди типов медиальных абзацев по предметно-логическому содержанию оказались следующие:

Таблица 6. Типы начальных абзацев научно-популярного текста по их предметно-логическому содержанию

№	Типы начальных абзацев	Код	Частота	
1	Констатация некоторого состояния главного объекта в зависимости от других главных и второстепенных объектов	M001	6	0,0645
2	Утверждение о состоянии главного объекта и констатация будущих действий автора	M002	1	0,0108
3	Констатация некоторой специфичности главного объекта в связи с другими главными и второстепенными объектами	M003	11	0,1183
4	Подчеркивание некот. специфич. глав. объекта и краткая констатация действ. главного субъекта	M004	13	0,1398
5	Констатация некотор. прошлого действия автора	M005	6	0,0645
6	Утверждение о сути некоторого главного объекта	M006	10	0,1075
7	Высказывание некоторого условия, связанного с будущим действием автора	M007	10	0,1075
8	Подчеркивание некоторой спец. главн. объекта и его влияние на глав. субъект и др. главный объект	M008	5	0,0538
9	Констатация некот-го действия автора, причины этого действия и оценка будущих действий автора	M009	5	0,0538
10	Высказывание предположения об особенностях глав. субъекта и усл. изменен. этих особенностей	M010	1	0,0108
11	Констатация некоторого простого утверждения автора, связанного с главным объектом	M011	9	0,0968
12	Описание некоторых особенностей главного объекта и действий прочих субъектов, связанных с этим главным объектом	M012	5	0,0538
13	Утверждение об особенностях прочих субъектов, связанных с главн. объектом	M013	1	0,0108
14	Констатация некоторого последующего действия автора, связанного с характер. главного объекта	M014	3	0,0323
15	Высказыван. предположения о сути глав. объекта	M015	2	0,0215
16	Оценка прошлого действия автора, констатация его будущего действия и постановка вопросов, связанных с главн. и второстепенными объектами	M016	1	0,0107
17	Подчеркивание особенностей второстепенных объектов и описание действий автора	M017	1	0,0107
18	Двустороннее описание особеннос. главн. объекта	M018	1	0,0107
19	Констатация последующих действий автора, связанных с главным объектом	M019	1	0,0107
20	Констатация прошлого действия автора, связанного с характеристикой главного объекта и констатация предстоящего действия автора	M020	1	0,0107
	ВСЕГО		93	1,0000

– «подчеркивание специфичности второстепенного объекта» (M111, f=0,1445);

– «констатация некоторого прошлого и настоящего действия автора» (M113, f=0,1048);

– «констатация некоторых условий, связанных с прочими второстепенными объектами» (М110, f=0,1048);

– «констатация прошлых действий автора и описание поведения главного субъекта» (М123, f=0,0652);

– «констатация состояния главного объекта» (М122, f=0,0537);

Таблица 7. Типы конечных абзацев научно-популярного текста по их предметно-логическому содержанию

№ п/п	Типы конечных абзацев	Код типов	Частота	
			F	f
1	Заключительное итоговое утверждение о сути главного объекта	М201	19	0,3333
2	Утвержд. автора об особен-тях главных объектов	М202	10	0,1754
3	Констатация прошлых действий автора, связанных с главным объектом	М203	3	0,0526
4	Постановка некоторых вопросов и констатация будущих действий автора	М204	2	0,0351
5	Описание будущих действий автора в связи с главным объектом	М205	1	0,0175
6	Заключительное утверждение, связан. с главным субъектом, с выходом на рисунок, графики и др.	М206	1	0,0175
7	Констатация предстоящих действия автора	М207	1	0,0175
8	Некоторая характеристика главного объекта с постановкой вопроса и оценка этого вопроса	М208	13	0,2283
9	Краткая двусторон. характери-ка главн. объекта	М209	4	0,0702
10	Высказывание предположений о будущем поведении главного и второстепенного объектов	М210	3	0,0526
	Всего		57	1,0000

Приведенное предметно-логическое содержание абзацев свидетельствует о том, что в основной части текста происходит сам процесс изучения главного объекта путем анализа действий (поведения) главных субъектов и некоторых действий (физических и умственных) автора, проводящего исследование.

Данные ***таблицы 7*** являются результатом анализа 57 конечных абзацев 47 научно-популярных казахских текстов. Здесь указаны 10 характерных для заключительной части текста типовых абзацев с кодовыми номерами от М201 по М210 и их частотами употребления. Подобное содержание конечных абзацев, указанных в таблице, вполне объяснимо, так как в заключении любого научного текста, как правило, формируются основные выводы, результаты проведенного исследования и намечаются действия автора (исследователя) по дальнейшему развитию исследования.

Итак, в организации содержания научного текста, независимо от его структурной части, существуют определенные общие закономерности, выражающиеся в употреблении в разных частях такого текста абзацев с одним и тем же предметно-логическим содержанием.

4.6 Особенности абзацев публицистического текста по их предметно–логическому содержанию

Аналогично изложенному в ***пункте 4.5***, мы выделили всевозможные типы абзацев по их предметно-логическому содержанию в 35 публицистических (газетных) текстах. Как было отмечено, это статьи из газеты «Қазақстан мұғалімі», посвященные одной тематике: «Алтыжасарлар мектепке барады». При анализе 25 начальных абзацев выявлено 11 различных по характеру типа для данной позиции абзацев, которые нами условно были обозначены кодовыми значениями от М301 по М311. Из них наиболее употребительными оказались:

– «высказывание автора о сути прочего объекта и констатация его состояния в связи с главным субъектом» (М310, f=0,16);

– «характеристика прочего объекта и высказывание предположениями о будущих действиях прочих субъектов» (М309, f=0,16);

– «описание природной или внешней обстановки, связанной с прочими субъектами и объектами» (М307, f=0,12) и т.д.

Перейдем далее к медиальным абзацам публицистического текста, т.е. к абзацам, где изложено основное содержание газетных статей. При анализе медиальных абзацев публицистического текста, можно утверждать, что выявленные типы в отношении предметно-логического содержания намного превосходят численность начальных и конечных абзацев.

Среди 175 медиальных абзацев публицистического текста нами зафиксирован 41 типовой вариант, относящийся к различным прдметно-логическим содержаниям. Им присвоены кодовые значения в интервале от М401 по М441. Самым частым из всех возможных случаев является тип М413 с относительной частотой 0,0914 (более 9% всех типов). Этот тип имеет формулировку «констатация автором некоторого факта, связанного с прочими субъектами и объектами». Далее следуют типы по мере убывания их частот.

В качестве конечных абзацев в публицистических текстах нами рассмотрено 45 абзацев, которые, на наш взгляд, являются заключительными абзацами в 35 газетных статьях. Анализируя ранее известным методом, использованным в классификации научно-популярного текста, мы обнаружили 11 типовых абзацев, относящихся к предметно-логическому содержанию, из которых более употребительны: «высказывание автором предположения, связанного с главным субъектом» (М509, f=0,2889); «развернутое рассуждение автора, связанное с главными и второстепенными объектами и субъектами» (М507, f=0,2002); «констатация заключительного утверждение автора о рассказанном» (М503, f=0,1778).

При сравнении между собой всех абзацев, использованных при создании публицистических (газетных) текстов можно отметить, что, в отличие от научно-популярных текстов, здесь существует определенная специфика в построении начальной, центральной и заключительной части

таких текстов. Наряду с этим, замечается и общность как в структурном, так и в логическом планах, которая составляет незначительный процент от числа всех абзацев, использованных при создании начальной, центральной и конечной части текстов.

4.7 Особенности абзацев художественного текста по их предметно-логическому содержанию

Предметно-логическое содержание абзацев художественных текстов проанализировано нами без учета позиций в тексте, т.е. мы не разделяли их на начальные, медиальные (центральные) и заключительные и рассматривали все 487 абзацев. При этом особое внимание было обращено на содержание каждого абзаца и учитывалась роль главных и второстепенных объектов и субъектов микроситуации. В результате такой трудоемкой процедуры нам удалось выделить 32 типовых по содержанию абзацев, формулировка и статистика самых употребительных среди них оказались следующие: «описание природы или внешней обстановки и констатация некоторого действия главного субъекта» (М601, f=0,1109); «описание действий главного субъекта, связанных с главным объектом и другим главным субъектом» (М605, f=0,0821); «описание природы и констатация действия или состояния автора» (М604, f=0,0636) и др.

Проанализировав статистические данные, отметим, что для художественного текста свойственны типы, относящиеся к описанию внешней среды, состояния или действия главных персонажей, а также характеристики главного субъекта, объекта и констатация будущих действий автора или прочих субъектов и т.д.

Таким образом, подход к абзацу как к минимальной семантико-синтаксической единице текста позволил нам выявить некоторые особенности трех рассматриваемых текстов по их предметно-логическому содержанию:

1. По предметно-логической классификации содержания научно-популярных текстов выделено: а) 20 типов начальных абзацев; б) 23 типа медиальных абзацев; в) 10 типов конечных абзацев.
2. По предметно-логической классификации содержания публицистических текстов выделено: а) 11 типов начальных абзацев; б) 41 типа медиальных абзацев; в) 11 типов конечных абзацев.
3. По предметно-логической классификации содержания художественных текстов выделено 32 типа абзацев без учета их позиции в структуре текста.

5 Основное статическое содержание текста и его формальное описание

5.1 Статика и динамика содержания текста

Статику текста следует понимать как слова «содержания», которые служат в тексте отражением некоторого множества предметов, явлений, фактов реальной действительности, а динамику – как синтаксис, отражающий отношения между элементами указанного множества. Пронизывающими весь текст на всей его протяженности и отражающими его смысловое единство, являются так называемые «опорные слова» или слова «содержания». Дифференцированный анализ «опорных слов» показывает, что в их структуре существует определенная иерархия, а именно часть таких «опорных слов» является «главными», «важными словами», всего содержания текста. Можно полагать, что эти слова, определяя главный предмет сообщения, представляют собой как бы свернутый замысел текста.

В абзацах отражаются те связи и отношения, которые зафиксированы у главных и второстепенных референтов текста с другими, менее значимыми элементами ситуации: различные типы действий и состояний ГОС и ВОС, различные места действий, время действий, цели, причины действий и состояний и т.п.

Второй тип составляющих текста – динамика текста, которая характеризует отношения между объектами реальной действительности, зафиксированные в конкретной ситуации, описанной в тексте.

5.2 Способы выделения опорных и ключевых слов текста

Проблема выделения основного и факультативного содержания текста является одной из важных задач в инженерной лингвистике. Из множества существующих способов выделения главных опорных слов (ГОС) наиболее перспективным нам представляется метод, в основе которого лежит идея о том, что такая наиболее значимая единица текста должна иметь, во-первых, наибольшую частоту встречаемости в тексте, во-вторых, встречаться в наибольшем количестве абзацев такого текста. Такими опорными словами, как правило, являются имена, отражающие главные действующие лица или главные предметы таких типов текста, как описание, повествование или рассуждение. Главные опорные слова в процессе развертывания содержания становятся своеобразным центром, вокруг которого формируются другие элементы как составляющие микроситуации. Такими составляющими обычно служат следующие классы слов: глаголы, прилагательные, наречия и т.д.

В зависимости от ситуации, главные референты (им соответствуют ГОС) могут быть взаимосвязаны с другими референтами, относительно важными для рассматриваемой ситуации, т.е. второстепенными опорными словами (ВОС). Естественно, что ВОС по частотности стоят ниже, чем ГОС. Слова текста, не относящиеся к ГОС и ВОС, мы назвали прочими словами, т.е. сокращенно ПС.

Отбор конкретных слов и включение их в различные типы сочетаний проводятся автором в зависимости от темы текста, т.е. «о ком (чем) будет текст». Под темой мы будем понимать совокупность определенных слов,

отражающая наиболее существенные составляющие описываемой в тексте ситуации и наиболее важные отношения между этими составляющими [Смит ДЖ., 1980].

К таким составляющим мы отнесем те слова текста, которые передают: 1) главные действующие лица; 2) главные объекты действия; 3) основные места действия; 4) времена действия.

Основная особенность таких последовательностей лексических единиц текста заключается в том, что они имеют в тексте одну и ту же предметную соотнесенность. В таком ряду могут присутствовать и полные словарные синонимы, и контекстуальные синонимы, и ассоциативные замены, и, наконец, замены местоименные [Леонтьева Н.Н., 1981]. Будем считать, что представленные четыре способа являются способами денотативного отождествления имен.

Большинство методов (анкетные, структурные, словарные, частотные, синтаксические, диалоговые), с помощью которых определяют «степень важности» слов для содержания текста, имеют один существенный недостаток: они не исходят из текста как единой связной единицы. Например, с помощью структурных методов выделяются опорные слова на основе их вхождения в заголовок, первое предложение или нескольких отдельных предложений. Частотные методы исходят из частоты встречаемости слов в тексте. При таком подходе по существу учитывается не смысл текста, а свойства плана его выражения [Зубов А.В., 1985].

5.3 Способ выделения опорных слов казахского текста

Как мы отмечали, основными критериями при выделении опорных слов текста являются абсолютная частота употребления слова (с учетом всех возможных замен) и количество абзацев, в которых встретилось слово, кроме этого, такой критерий не должен зависеть от общего числа слов в тексте.

Наиболее удобной для нас представляется несколько измененная формула коэффициента статистической устойчивости термина [Марусенко М.А., 1983]:

$$K\text{важ} = \frac{\mathrm{Fm}}{\mathrm{Nn}} \ . \qquad (1)$$

где F – абсолютная частота слова в тексте (в нее входит суммарная частота всех типов синонимов этого слова и местоименных замен); m – число абзацев, в которых встретилось слово; N – общие число слов в тексте; n – общее число абзацев в тексте.

Назовем этот критерий *Кважс* **коэффициентом важности слова** и определим для слов всех анализируемых текстов его критические значения K^1*важ* и K^2*важ*, позволяющие формальным способом отделить в массе слов конкретного текста соответственно главные и второстепенные опорные слова.

Эксперименты по выделению опорных слов дают нам возможность предложить на основе формулы (1) составленные экспериментальным путем две формулы-неравенства (2) и (3):

$$\frac{\left(\frac{n}{m_1}+1\right)^2}{N \cdot n} \le \text{К}^1\text{важ} \le 1 \quad \textbf{(2);} \qquad \frac{\left(\frac{n}{m_2}+1\right)^2}{N \cdot n} \le \text{К}^2\text{важ} < \frac{\left(\frac{n}{m_1}+1\right)^2}{N \cdot n} \quad (3).$$

Здесь неравенство (2) служит для определения границ ГОС, а (3) – для ВОС. Значения m_1 и m_2 будут зависеть от интервалов, в которые попадет значение *n* – общее число абзацев в тексте. Например, для ***К¹*** важ, если общее число абзацев *n* равно или больше 100, то m_1=42, а если *n* больше 60, но менее 100, то m_1=30 и т.д. Полный перечень значений для m_1 и m_2, найденных экспериментальным путем, и соответствующие им формулы неравенств для ***К¹*****важ** и ***К²*****важ** приведены в книге «Основные принципы формализации содержания казахского текста» [Жубанов А.К., 2002, с.151-153].

5.4 Выделение опорных слов в трех анализируемых типах текстов

Из распределительных алфавитно-частотных словарей были получены потенциальные словари, которые являются базовыми словарями для определения ГОС, ВОС и прочих слов рассматриваемых текстов. Для отдельного типа текстов и входящих в него минитекстов по известному общему числу слов в тексте *N*, общему числу абзацев в каждом минитексте ***n*** заранее были вычислены пределы изменения *К¹*важ и *К²*важ. Здесь эти данные представлены только для художественного текста в ***таблице 8*** и фрагмента (отрывка) потенциального словаря в виде ***таблицы 9***, рассматривающей XUD-03 (художественный) и слова, относящиеся к классу существительных.

Во всех потенциальных распределительных алфавитно-частотных словарях для каждого слова нами вычислены значения $К_{важ}$ и сравнением их со значениями *К¹*важ и *К²*важ (согласно типу текста) была установлена принадлежность слов к ГОС, ВОС и ПС. Эти сведения записаны в последней графе ***табл. 9***. В качестве примера рассмотрим слово «***ауыл»*** из потенциального словаря, составленного по тексту XUD-03, *N=361, n=11* (***табл. 8, 9***).

Согласно формуле (1), $К_{важ}=Fx\ m/Nx\ n$; $Fx\ m=5x4=20$, $Nxn=361x11=3971$, то $К_{важ}=20/3971=0{,}00503$. Значения $К^1_{важ}$ и $К^2_{важ}$ вычисляют по формулам (2) и (3) и соответственно равны 0,0026 и 0,0008.

Таким образом, значение $К_{важ}=0{,}00503$ больше, чем $К^1_{важ}=0{,}0026$, но меньше чем 1, а это удовлетворяет условию, когда слово из текста XUD-03 «***ауыл»*** должно принадлежать к ГОС.

Таблица 8. Пределы изменения значений $К^1_{важ.}$ и $К^2_{важ.}$ для художественного типа казахских текстов

№ п/п	Имя текста (XUD-ij)	Число слов N	Число абзацев n	N•n	$К^1_{важ}$: от $К^1$ до 1 $К^1$	$К^2_{важ}$: от $К^1$ до $К^2$ $К^2$
1	XUD-01	274	4	1096	0,0082	0,0036
2	XUD-02	370	13	4810	0,0027	0,0008
3	XUD-03	361	11	3971	0,0026	0,0008
4	XUD-04	2432	74	179968	0,0002	0,00005
5	XUD-05	753	43	32379	0,0004	0,00015
6	XUD-06	1007	54	54378	0,0003	0,00011
7	XUD-07	347	9	3123	0,0025	0,00086
8	XUD-08	150	4	600	0,0150	0,00670
9	XUD-09	473	17	8041	0,00073	0,00031
10	XUD-10	1121	33	36993	0,00025	0,00010
11	XUD-11	988	33	32604	0,00029	0,00011
12	XUD-12	218	15	3270	0,00155	0,00069
13	XUD-13	150	8	1200	0,00750	0,00330
14	XUD-14	577	24	13878	0,00065	0,00023
15	XUD-15	1144	37	42328	0,00026	0,000097
16	XUD-16	1233	37	45621	0,00024	0,00009
17	XUD-17	908	18	16344	0,00038	0,000157
18	XUD-18	115	7	805	0,00940	0,00437
19	XUD-19	173	9	1557	0,00504	0,00173
20	XUD-20	572	37	21164	0,00052	0,00019

Таблица 9. Потенциальный алфавитно-частотный распределительный словарь художественного текста XUD-03 *(фрагмент)*

XUD-03: $К^1_{важ} \geq 0,0026$; $К^2_{важ} \geq 0,0008$					
№ п/п	**Слово**	**Абсолютная частота $F \geq 2$**	**Число абзацев m**	**$К_{важ}$**	**ГОС, ВОС, ПС**
1	Ақан/ат	5	3	0,00377	ГОС
2	Ақмола/жа	2	2	0,00100	ВОС
3	ара/зт	2	2	0,00100	ВОС
4	ауыл/зт	5	4	0,00503	ГОС
5	банды/зт	2	2	0,00100	ВОС
6	ел/зт	2	2	0,00100	ВОС
7	ет/зт	2	2	0,00100	ВОС
8	жолдас/зт	3	2	0,00151	ВОС
9	жігіт/зт	2	2	0,00100	ВОС
10	комитет/зт	3	1	0,00075	ПС
11	Кудря/ат	6	3	0,00453	ГОС

5.5 Построение таблицы основного статического содержания текста

Известно, что при задании содержания текста должно быть указано *кто (*или *что)* совершает события, над *чем (кем)* совершаются они, *где и в*

какое время, т.е. это есть не что иное, как отражение основных категорий диалектического материализма – материи, движения, времени и пространства. Поэтому среди опорных слов текста можно особо выделить и записать в виде «Таблицы основного статического содержания» такие типы, которые относятся к словам-субъектам, словам-объектам, а также словам, означающим место и время прохождения события в зависимости от их десигнативной отнесенности. В ***таблице 10*** конкретизированы типы опорных слов и каждому типу присвоен буквенно-цифровой код. Цифровые индексы могут изменяться от 01 до 99 в зависимости от количества слов, отнесенных к данному типу.

Таблица 10. **Таблица основного статического содержания (ТОСС)**

Группа опорных слов	Тип опорного слова	Код
Главные опорные слова (ГОС)	Слова-субъекты	Cij
	Слова-объекты	Kij
	Слова, обозначающие место	Mij
	Слова, обозначающие время	Tij
Второстепенные опорные слова (ВОС)	Слова-субъекты	Sij
	Слова-объекты	Gij
	Слова, обозначающие место	Pij
	Слова, обозначающие время	Bij

По потенциальному словарю каждого текста можно выделить главные и второстепенные опорные слова, и согласно содержанию рассматриваемого текста, отнести их к соответствующим типам, указанным в ***таблице 10***.

Таблица 11. Таблица основного статического содержания (ТОСС по PUB-06)

Конкретное значение опорных слов	**Группа опорных слов**			
	Главные опорные слова	**Код ГОС**	**Второстепенные опорные слова**	**Код ВОС**
Субъекты	*оқушы*	C11	*пән*	S11
	–	–	*мұғалім (пән мғалімі)*	S12
Объекты	*тіл (ана тілі)*	K11	*тақырып (еркін)*	G11
	әдебиет (тіл мен әдебиет)	K12	*мүше (секция мүшесі)*	G12
	апта (апта күндері)	K13	*қазына (асыл қазына)*	G13
	апталық	K14	*сабақ*	G14
	пән (пән апталығы)	K15	*білім (білім тереңдету)*	G15
Место Время	*мектеп*	M11	*класс (7 ª)*	P11
	(алғашқы күн)	T11/01	–	–
	(екінші күн)	T11/02	–	–
	(үшінші күн)	T11/03	–	–
	(төртінші күн)	T11/04	–	–
	(бесінші күн)	T11/05	–	–
	(сенбі күні)	T11/06	–	–
	кеш (қорытынды кеш)	T12	–	–

В качестве примера рассмотрим публицистический текст PUB-06 и покажем, как конкретизируются опорные слова в виде ***таблицы 11*** «Основного статического содержания».

Слова, отражающие элементарные составляющие микроситуации (ЭСМС), в определенной степени достраивают «костяк» содержания текста, заданный таблицей основного статического содержания, но они не завершают его полное построение. Это объясняется тем, что, кроме опорных слов и слов-конкретизаторов, в любом тексте существует еще одна группа слов, которую называют словами-наполнителями. Они «наполняют» полученное содержание качественными признаками, т.е. указывают качества предметов, названных именами существительными, признаки действий, определяемые в тексте глаголами и причастиями [Зубов А.В., 1985].

Экспериментальные данные, представляющие собой слова-конкретизаторы и слова-наполнители для казахских текстов, в этой работе не рассматриваются. Отметим только то, что исследования по этому вопросу для различных функциональных стилей казахского языка требуют дальнейшего изучения.

5.6 Средства связи в предложении, между предложениями в абзаце и между абзацами

Виды и средства синтаксической связи в словосочетаниях, а также в простых и сложных предложениях казахского языка достаточно хорошо изучены и освещены в академических грамматиках казахского языка. В работе Ж.Т.Кайшыгуловой «Қазақ мәтінінің күрделі фразалық тұтасым//абзац деңгейінде мүшеленуі» перечислены средства связи отдельных предложений, которые создают целостность текста: «*Мәтін құрылымына кіретін сөйлемдердің байланыстылығын білдіретін формалды құралдар, қайталанатын сөздер, жіктеу (III жақ) және сілтеу есімдіктері, синоним, антонимдер, түбірлес сөздер, жалғаулықты шылаулар, демеуліктер т.б.*». Эти средства связи могут также встречаться внутри простых и сложных предложений.

К формализуемым средствам связи предложений письменной речи нами отнесены следующие: 1) связи через формы слов; 2) связи с помощью различного рода союзов и союзных слов *(шылау сөздер);* 3) связи через порядок слов в предложении.

Из перечисленных средств связи мы остановимся только на тех, которые осуществляются с помощью различных союзов и союзных слов.

Задание каждого союза или союзного слова отдельным кодом связано с тем, что практически невозможно найти правила, которые бы определяли их выбор в предложении. Их выбор определяется субъективными особенностями автора текста.

На основании данных, полученных в результате анализа средств связи трех видов рассмотренных нами текстов, и, учитывая некоторые идеи ученых [Берзон В.Е., Зубов А.В., 1977], мы можем предложить следующую

классификацию формальных средств связи между предложениями в абзаце: 1) соединительные элементы; 2) повторение лексических форм; 3) замещения; 4) ассоциативные синонимические связи; 5) ассоциативные антонимические связи; 6) грамматические средства связи.

Перечислим случаи, которые чаще других являются соединительными элементами связи между предложениями в абзаце:

1. Связи с помощью союзов и союзных слов: *оның үстіне, тағы да, неге десең, сонымен бірге, оның есесіне, оның орнына, солай бола тұрса да, әрине, шынын айтқанда, сөйтіп, сонымен, сондықтан, сол себепті, сол үшін, соның нәтижесінде;* разделительные союзы: *я, не, немесе, не болмаса, біресе, яки,* условный союз *егер* и т.д.

2. Связи с помощью слов, которые выступают в роли союзов – слова-местоимения: *қандай, сондай, қанша, сонша, кім, сол, не, қайсы, қайда, қайдан, содан, қалай* и т. д.

3. Связи с помощью различного рода связок – логико-смысловых скрепов: а) результат, итог: *қорытсақ, қорыта айтқанда, қорытындысы, нәтижесінде, ақырында, соңында, соң, сосын, айта келе, айта берсек, дәл осы, тек қана, осының барлығы* и т.д.; б) сопутствующие обстоятельства, дополнения: *оның үстіне, одан басқа, ол аз болса, сонымен бірге, онан әрі, әрі қарай, ары, мұның үстіне, онан кейін, одан әрі, одан басқа, осыдан басқа, бола тұра, солай бола тұра, соған ұқсас, не үшін, сол үшін, дәл солай* и т.д.; в) возражения: *бірақ, бірақ та, алайда, дегенмен, солай бола тұрса да, әйтсе де, сөйтсе де, әйткенмен, керісінше, өйткенмен, сонда да* и т. д.; г) оценки достоверности сообщения, подтверждения, положительной аргументации: *дұрысында, шындығында, дұрысы, шыны, шынында, расында, асыра айтса, кем дегенде, сүйте тұра, солай бола тұра, табиғи, әрине, орынды, өте орынды, мәні мынада, одан да артық* и т. д.; д) причинно-следственные связи: *туындайды, сондықтан, демек, солай болғасын, олай болса, солай болғандықтан, солай болған соң, сонда, әсерінен, нәтижесінде, оның орнына, солай бола тұра, осы себептен, дәл сол себептен* и т. д.; е) обстоятельственные отношения: *бұл жерде, мұнда, мында, содан бері, осыған дейін, онда, енді, осында, сонымен бірге, сонда да, барлығына тән, барлығына ортақ* и т.д.; ж) субъективное отношение автора к высказанному в тексте: *тән, жиі, сирек, оңай, жеңіл, қиын, көруге оңай, айтуға оңай, жеңіл мінез, байқауға қиын, шындық, айта кету қажет, күмән тудырады, күмәнсыз* и т.п.

К группе логико-смысловых скрепов можно отнести также следующие: з) перифразы: *басқа сөзбен (айтсақ), басқаша айтсақ, өзге сөзбен (айтқанда), шындығында, шынайы айтса, сөз болып отырған* и др.; и) перечислительные конструкции: *біріншіден, екіншіден, бір жағынан (алғанда, қарастырғанда), басқа жағынан (алсақ, келсек), бірінші, екінші, үшінші* и т.д.; к) слова, предшествующие приведению примера, факта или некоторой аргументации: *соның ішінде, сонымен қатар, сонымен бірге, сондай-ақ, мысал ретінде, мысалы, сонымен, «сонымен, мысал»* и т. д.

Среди таких средств связи, как «повторение лексических форм», может участвовать повторение слов и словосочетаний с отсылкой, например *сол тәрізді, сол сияқты, берілгендей, қарастырылғандай, бұл, одан басқа, одан бөтен, тіпті, басқа, қалғандары* и т.д.

Все эти и другие соединительные элементы должны быть специально заменены условно принятыми кодами для программы порождения казахского текста.

Рассматривая формальные средства связи между абзацами в пределах некоторого текста, многие исследователи отмечают, что такие средства в целом совпадают с теми формальными средствами, которые соединяют предложения в абзаце [Котюрова М.П., 1979; Ляховицкая Л.М., 1973]. Здесь достаточно отметить, что частое употребление тех или иных средств связи зависит, в основном, от типа текстов (повествование, описание, рассуждение).

Для рассматриваемых нами типов текстов весьма характерными средствами межабзацных связей являются ассоциативные синонимичные связи, которые могут иметь следующие разновидности:

1. Связь через общеупотребительные или словарные синонимы: *ағайын – туыс – жақын, адам – кісі, айқын – анық – нақты, айырмашылық – өзгешелік – ерекшелік, аласа – тапал – қысқа – кішкене – мықыр – қортық, аман – есен – сау – саулық – сәлемет, ас – тамақ – дәм – тағам, әдіс – тәсіл – амал – жол* и т. д.

2. Связь через различного рода контекстуальные синонимы: а) через понятия, объединяемый ассоциативными синонимами типа «род» – «вид»: *колхозшы – малшы, егінші, ғалым – тілші, әдебиетші, тарихшы ..., ... ғылым – тіл, математика, медицина, философия, ..., дәрігер – хирург, стоматолог, терапевт, ...* и т.д.; б) связь через ассоциативные слова типа «целое» – «часть» и «часть» – «целое»: *адам – қол; жапырақ – өсімдік; үй – бөлме, терезе, есік; ағаш – табиғат* и т. д.; в) связь через ассоциативные слова, объединенные по принципу разной степени одного и того же признака или интенсивности обозначения действия: *жылы – ыстық – күйіп тұр, жанып тұр – суық – салқын; баяу – тез – жылдам – шапшаң – жүйрік (жүріс); сұйық – қою – қатты (зат)* и т. д.; г) связь через ассоциативные слова, объединенные по принципу ассоциативного переноса понятии по таким категориям, как «причина» – «следствие», «предмет» – «материал» – «качество», «действие» – «объект» – «деятель”: *суық – тоңу (мұздау, қату); ыстық – еру – терлеу; шелек – темір – ыдыс; үй – ағаш – тас – шым – көркем – сұлу – биік; ату – мылтық; ұшу – құс – ұшақ – ұшқыш* и т. д.; д) связь через ассоциативные слова, объединенные любой другой общностью коннотативных значений: *ауру – дәрігер – аурухана, қасқыр – ұлу, ит – үру, құс – ұшу* и т.д.

Ассоциативные синонимические связи между абзацами являются основой статического содержания текста и формируют главные и второстепенные опорные слова таблицы основного содержания текста.

Существенное отличие связей между предложениями в абзаце от связей абзацев в тексте заключается в том, что первые, как правило, фиксируются в соседних предложениях абзаца, в то время как абзацы могут быть связаны дистантными связями. Мы принимаем случай, когда формальные средства связи между абзацами в тексте совпадают с такими же средствами для связи предложений в абзаце.

5.7 О семантико-синтаксическом языке для записи формул предложения, абзаца и текста

Изучая проблему организации текстов из элементов разных уровней, мы должны найти способ фиксации в абзацах реальных текстов всех его компонентов. Возникает необходимость создания искусственного языка, который мог бы служить некоторой моделью естественного языка (ЕЯ).

Мы остановимся только на том случае, когда моделирование ЕЯ сводится к статистическому выбору из текстов наиболее употребительной лексики и грамматических правил, относящихся к узкой предметной области.

В нашем представлении такой искусственный язык (ИЯ) должен иметь возможность описать абзац текста как некоторую семантико-синтаксическую единицу, учитывающую статический и динамический аспекты, т.е. ИЯ может строить некоторую формулу абзаца. Если любой абзац состоит из цепочки взаимосвязанных предложений, то искомый язык должен содержать правила, позволяющие учитывать семантические и синтаксические отношения между членами предложения, т.е. он должен строить семантико-синтаксическую формулу предложения. В нем также должны участвовать элементы, отражающие логико-смысловые связи между предложениями одного абзаца и между абзацами.

И наконец, для описания содержания всего текста через содержание отдельных абзацев в ИЯ (искуственный язяк) должны быть средства для создания из формул абзацев некоторые формулы текста.

Следуя положениям, разработанным А.В.Зубовым [Зубов А.В. 1987, 1988], мы для описания содержания текста в языке СЕМСИНТ (семантико-синтаксический) также включили следующие составляющие: 1) алфавит языка СЕМСИНТ; 2) набор элементов для записи семантических отношений между членами предложения; 3) набор элементов для записи синтаксических отношений между членами предложений; 4) набор элементов для фиксации логико-семантических отношений между предложениями одного абзаца; 5) набор элементов для связи в абзаце статической и динамической составляющих текста; 6) семантико-синтаксическую формулу абзаца (СЕСФА); 7) набор элементов для фиксации связей между абзацами одного текста; 8) семантико-синтаксическую формулу текста (СЕСФТА).

При анализе предложений в абзацах исследуемых текстов можно выделить семантические роли участников ситуации (или семантические функции), отображенные в предложении. Кроме того, для написания семантико-синтаксической формулы предложения нам необходимы

Таблица 12. Семантические функции, используемые в языке СЕМСИНТ и их синтаксические значения

№ п/п	Семантическая функция	Содержание семантической функции. Характеризует аргументную группу как отражение в предложении в следующих событиях:	Код
1	Субъект	Активного одушев-го инициатора события (действия, состояния, процесса) вместе с признаками, выражаемого глаг-м предложения	AAG
2	Предмет	Активного неодушевленного инициатора события – предмета вместе с его признаками	AH1
3	Понятие (общее)	Активного неодушевленного инициатора события – общего понятия вместе с признаками	AS2
4	Понятие (научное)	Активного неодушевленного события –научного понятия вместе с его признаками	AS1
5	Процесс	Активного неодушевленного инициатора события – процесса с его признаками	AS3
6	Деятель (неодушевленный)	Активного неодушевленного инициатора события – природного или стихийного явления или предмета вместе с их признаками	AEL
7	Объект	Одушевленного объекта некоторого события вместе с его признаками	AO
8	Получатель	Одушев-го адресата, получателя некоторого события, выражаемого глаголом предложения	AB1
9	Адресат	Неодушев-го адресата некоторого события	AB2
10	Объект	Неодушев-го объекта некоторого события, выражаемого глаголом предложения	AH
11	Место	Места совершения некоторого события, выражаемого глаголом предложения	ALK
12	Средство	Средства выполнения события, выражаемого глаголом предложения	AMD
13	Инструмент	Орудия или инструмента выполнения события, выраженного гл-м предложения	AIN
14	Состав	Состава или содержимого какого-либо предмета, выражаемого в предложении др. аргументной группой	AKM
15	Определитель	Некот-го св-ва какого-либо предмета, выраженного в предл-и своей аргумент-й группой	AD1
16	Время	Времени осуществления события, выражаемого глаголом предложения	ATM
17	Способ	Способа реализации события, выраженного глаголом предложения	AAD
18	Условие	Условия осуществления события, выраженного глаголом предложения	AIF
19	Причина	Причины осуществления события, выраженного глаголом предложения	ACS
20	Цель	Цели выполнения события, выраженного глаголом предложения	AAM
21	Вставка	Некоторой вспомогательной информации (примеры, таблицы, графики, рисунки)	IN_i i=1,2...

специальные коды, по значениям которых узнаются содержание каждой семантической функции, а также их синтаксические значения в СЕМСИНТ (***таблица 12***).

Искусственный язык (ИЯ), принятый на основе ЕЯ, разработанный А.В.Зубовым и названный семантико-синтаксическим языком, с некоторыми дополнениями, специфичными для казахского языка вполне может удовлетворять требованиям, которые ставятся при формализованной записи содержания казахского текста.

Итак, приведенные сведения об алфавите языка СЕМСИНТ, а также некоторые формальные средства связи слов в предложении и между предложениями абзаца будут применяться в дальнейших наших разработках по этой проблеме.

Семантико-синтаксическая формула предложения (СЕСФП), которая может быть представлена в виде C_i («смысл», где i=1,2,3,...) строится с помощью 21 семантической функции (табл.12), разделенной по значению требуемых ими поверхностных падежей, а также кодов союзов, союзных слов и других положений. В формуле предложения будут использованы коды семантических подклассов казахских существительных, глаголов, прилагательных, причастий, деепричастий, наречий и местоимений.

На основании формул предложений строится уже семантико-синтаксическая формула абзаца (СЕСФА) как детерминированная последовательность формул предложений C_i (СЕСФП). Если СЕСФА представить в виде M_i («микротема»), то имеем:

$$M_i = C_1 \,\&\, C_2 \,\&\, C_3 \,\&\, \ldots \,\&\, C_n \,.$$

С учетом изложенного и того, что текст состоит из ограниченного количества абзацев, семантико-синтаксическую формулу текста (СЕСФТ) в свернутом виде можно представить как последовательность СЕСФА:

$$Ti=M_1(D_1/B_1)\&M_2(D_2/B_2)\&M_3(D_3/B_3) \ldots \&M_n(D_n/B_n),$$

где Ti – тема всего текста; Mi – микротемы, заключенные в абзацах i; Di – детерминированные правила, связывающие абзац «i» с абзацем «(i-1)»; Bi - вероятностные факторы, связывающие абзац «i» с абзацем «(i-1)».

Практическое написание формул предложений, абзаца и конкретного казахского текста вполне возможно, если воспользоваться материалами нашего исследования. А для их реализации на компьютере необходимы дополнительные исследования специалистов по инженерной лингвистике.

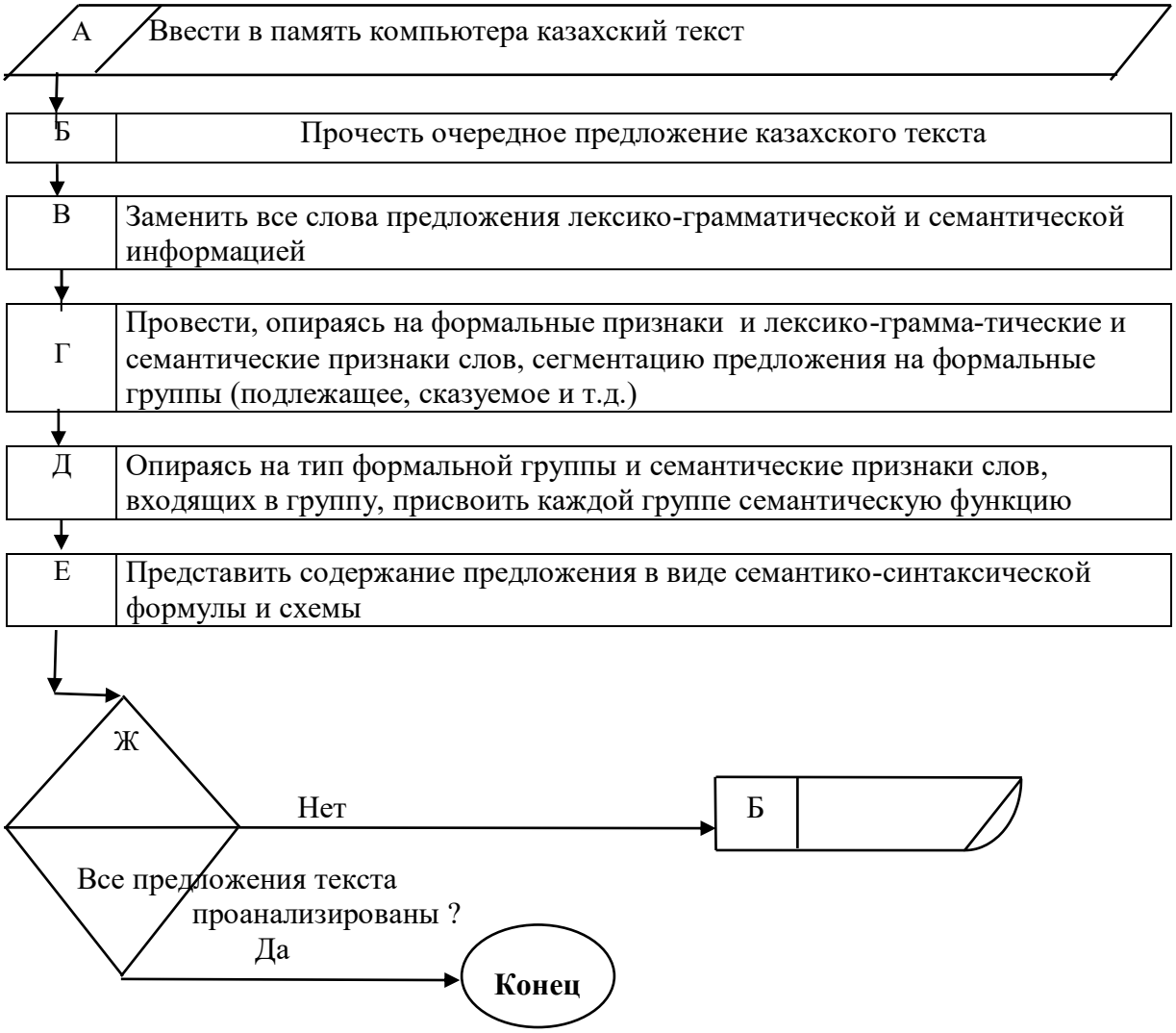
В заключение предлагаем принципиальный алгоритм перехода от естественного казахского текста к его формальному представлению, сопровождая его конкретным казахским предложением (см. блок-схему).

<u>Блок А.</u> Пусть вводится в память компьютера следующий текст: *«Үлкен жүк тасыйтын машина кешкі сағат алтыда теміржол көпірінен өтті. Ол Алматыға азық-түлік жеткізеді».* («*Большая грузовая машина в шесть часов вечера переехала железнодорожный мост. Она доставит продовольствие в Алматы*»).

<u>Блок Б</u>. Читаем очередное предложение: *«Үлкен жүк тасыйтын машина кешкі сағат алтыда теміржол көпірінен өтті».*

<u>Блок В</u>. Заменяем все слова этого предложения лексико-грамматической и семантической информацией из автоматического словаря казахского языка (АСКЯ). Условимся указать только некоторые, например, принадлежность данного слова к определенному классу – существительное (ЗТ), глагол (ЕТ) и т.д., а также код семантического подкласса:

Блок-схема. **Принципиальный алгоритм перехода от естественного казахского текста к его формальному представлению**



1) рассмотрим первое слово предложения: ***үлкен*** – прилаг. – (СН), качествен. – (ССijk) и по семантике – означает объем предмета, код которого «СС23», им. падеж, т.е. код падежа – (АС) от «*атау септік*». Таким образом, слово ***үлкен*** будет: (СН)(АС)(СС23);

2) ***жүк*** – существ. (ЗТ), им. падеж (АС), нарицат. (ЖЕ), неодушев. (ЖЗ), сем. кл. – предметы-артефакты, используемые человеком – ЗТ315 и др.; вместо слова ***жүк:*** (ЗТ)(АС)(ЗТ315);

3) ***тасыйтын*** – гл. (ЕТ) в форме причастия, прош. вр. (ӨШ), по сем. подкл. как причастие означает физическое действие, свойственое живому и неживому в природе – ЕШ081 и др., т.е. *тасыйтын* – (ЕТ)(ӨШ)(ЕШ081);

4) ***машина*** – существ. (ЗТ), ед. число (АС), нарицат. (ЖЕ), неодушев. (ЖЗ), сем. подкл. ЗТ308 и подкл.: объект, использ. для передвиж. чел-м; итак, слово ***машина*** – (ЗТ)(АС)(ЗТ308);

5) ***кешкі*** – прилаг. (СН), им. падеж (АС); сем. подкл. как относ. прилаг., характеризует объект по его отношению ко времени (ҚС031). Таким образом слово «*кешкі*» – (СН)(АС)(ҚС031);

6) ***сағат*** – существ. (ЗТ), им. пад. (АС), нариц. (ЖЕ), как сущес., означ. времен. понятие – (ЗТ601); то слово: (ЗТ)(АС)(ЗТ601);

7) ***алтыда*** – числительное (СА), местный падеж (ЖС), количественное (ЕК), т.е. все слово – (СА)(ЖС)(ЕК);

8) ***теміржол*** – сущ.. (ЗТ), ед. чис. (АС), нариц. (ЖЕ), по сем.: «понятия, связан. с устройст. отдельного механического объекта» и его код ЗТ117. Итак, слово «*теміржол*» – (ЗТ)(АС)(ЗТ117);

9) ***көпірінен*** – сущес. (ЗТ), исход. пад. (ШС), по сем. – ЗТ117, а само слово: (ЗТ)(ШС)(ЗТ117);

10) ***өтті*** – гл. (ЕТ), прошед. вр. (ӨШ), по сем. подкл. – глагол со знач. каузации движения: «обозначающие преодоление преграды в движении» – ЕТ941, то слово – (ЕТ)(ӨШ)(ЕТ941).

Теперь, заменив каждое слово в этом предложении соответствующими кодами, будем иметь: (СН)(АС)(СС23) (ЗТ)(АС)(ЗТ315) (ЕТ)(ӨШ)(ЕШ081)

(ЗТ)(АС)(ЗТ308) (СН)(АС)(ҚС031) (ЗТ)(АС)(ЗТ601) (СА)(ЖС)(ЕК) (ЗТ)(АС) (ЗТ117) (ЗТ)(ШС)(ЗТ117)(ЕТ)(ӨШ)(ЕТ941).

Принятые в этой записи наборы кодов абсолютно условные, которые зависят от того, как был составлен автоматический словарь казахского языка (АСКЯ).

<u>Блок Г.</u> Проведем сегментацию предложения на группы или на так называемые аргументные группы. Рассматриваемое предложение можно разбить на следующие четыре аргументные группы: 1) *үлкен жүк таситын машина* – именная группа; 2) *кешкі сағат алтыда* – обстоятельственная группа; 3) *теміржол көпірінен* – именная группа; 4) *өтті* – глагольная группа.

<u>Блок Д.</u> Зная тип группы (именная, глагольная, обстоятельственная) и падежи существительных, входящих в группы и их семантические характеристики, присваиваем этим группам наименования семантических функций (субъект, предмет, понятия и т.д.):

1) именная группа: «*үлкен жүк таситын машина*» – имя существ. в имен. пад., неодушев.; по значению семан. функции – «Предмет» с признаками: *үлкен* и *жүк таситын;* по данным ***табл. 13*** код «Предмет» – АН1. Теперь запишем предмет АН1, отбирая индекс класса и семантический подкласс:

АН1<(СН)(АС)(СС23)*(ЗТ)(АС)(ЗТ315)*(ЕТ)(ӨШ)(ЕШ081)*
(ЗТ)(АС)(ЗТ308)> ;

2) обстоятельственная группа: «*кешкі сағат алтыда*», аналогично по ***табл. 13*** код функции – АТМ, который равен:

АТМ<(СН)(АС)(ҚС031) * (ЗТ)(АС)(ЗТ601) * (СА)(ЖС)(ЕК)>;

3) именная группа: «*теміржол кӨпірінен»* – неодушевленный объект некоторого события, который имеет код АН, тогда:

АН<(ЗТ)(АС)(ЗТ117) * (ЗТ)(ШС)(ЗТ117)>;

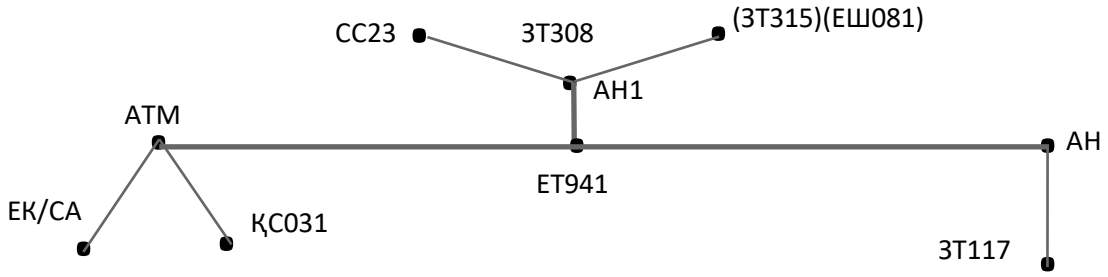
4) глагольная группа: *«өтті»* – (ЕТ)(ӨШ)(ЕТ941).

<u>Блок Е.</u> В итоге все предложение записывается в виде семантико-синтаксической формулы, понятной компьютеру:

C_1=АН1<(СН)(АС)(СС23)*(ЗТ)(АС)(ЗТ315)**(ЕТ)(ӨШ)(ЕШ081)*
(ЗТ)(АС)(ЗТ308)>+АТМ<(СН)(АС)(ҚС031)(ЗТ)(АС)(ЗТ601)*
*(СА)(ЖС)(ЕК)>+АН<(ЗТ)(АС)(ЗТ117)**(ЗТ)(ШС)(ЗТ117)>+
+(ЕТ)(ӨШ)(ЕТ941).

Представим теперь содержание предложения в виде следующей схемы, которую можно прочитать так: нечто (АН1) неодушевленное (ЗТ308), имеющее признак (СС23) и с определенным предназначением (ЕШ081), совершает физическое действие (ЕТ941) во время АТМ, связанное с некоторым объектом (АН), обладающим свойством (СН/ЗТ117).

Таким образом, можно было бы проанализировать второе предложение минитекста и перейти от естественной его формы к формальному виду, так же, как и в случае с первым предложением. Указанное правило распространяется и для любого казахского текста.



ОСНОВНЫЕ ВЫВОДЫ

Постановка вопроса по поиску основных принципов формализации содержания текста как общепринятой единицы общения людей, относительно казахского языка, ставится на повестку дня впервые. Исходя из этого, в диссертационной работе в рамках единой теории текстообразования последовательно рассматриваются проблемы лингвистики текста – основного вопроса современной лингвистики. Решаются задачи формального представления словаря казахского языка и абзаца, как основной семантически значимой текстовой единицы, а также поиска тех правил, понятных компьютеру, которые объединяют эти единицы в связный текст.

Важным моментом нашего исследования явилась семантическая классификация знаменательных классов слов казахского языка и их кодификация, которая позволяет передавать некоторую жизненную ситуацию, описываемую в тексте.

Проанализировав высказывания ученых разных направлений и специальностей о роли вероятности в общей теории познания, процессах

функционирования языка и организации различных текстов, мы можем сделать следующие выводы:

– все связи в реальном мире носят объективный характер, что проявляется в массе однородных явлений;

– язык, являясь саморегулирующей системой, при порождении речи функционирует, используя как детерминированные правила, так и случайные факторы;

– текст, как результат речевой деятельности, также содержит детерминированные и вероятностные составляющие;

– абзацы, как «кванты» содержания и «застывшие» синтаксические шаблоны, организуются в единое целое по вероятностно-алгоритмическому принципу.

Анализ конкретного материала, содержащего 102 минитекста в трех жанровых разновидностях, позволил нам предложить формальный критерий – коэффициент важности $K_{важ}$, позволяющий отделить опорные слова казахского текста от остальных его слов. При этом, удалось дифференцировать $K_{важ}$ на два критерия – $K^1_{важ}$ и $K^2_{важ}$, позволяющие выделить главные и второстепенные опорные слова. Разработаны конкретные числовые критерии выделения опорных слов для текстов разной длины – от 1 до 100 и более абзацев. Выделенные по указанному критерию опорные слова казахского текста (ГОС, ВОС), сыграли определенную роль в разработке построения «Таблицы основного статического содержания», являющейся своеобразным замыслом текста, который служит инвариантом, остающимся неизменным как для автора текста, так и для любого его получателя.

На основе анализа (осмысления) содержания трех разновидностей казахских текстов нами составлена специальная таблица, отражающая семантические роли участников ситуации. В таблице приведена 21 семантическая функция участников ситуации и раскрыто их содержание. Каждая семантическая функция имеет условный индивидуальный код (***табл.12***).

Разработан принципиальный алгоритм перехода от естественного казахского текста к его формальному представлению, которое создает широкие возможности для эксперимента по реализации задачи формализованной записи содержания казахского текста.

При изучении процедуры вероятностно-алгоритмического порождения казахского текста, нами была сформулирована основная идея и уточнены следующие гипотезы:

1. Текст является отражением некоторой жизненной ситуации, воспринимаемой с точки зрения конкретного автора. Минимальным смысловым «куском» или «квантом содержания» является абзац, показывающий некоторую микроситуацию.

2. Основой или темой текста служит некоторое число опорных слов, которые отражают главные субъекты, объекты, основное место и время событии, описанных в тексте.

3. Динамика текста понимается нами как «приведение в действие» его опорных слов, т.е. последовательная цепь их отношений между собой и с другими словами в разных частях текста.

5. Текст с заданной таблицей основного статистического содержания (ТОСС) представляется в виде семантико-синтаксической формулы текста (СЕСФТ), состоящего из последовательности СЕСФА, типичной для данного автора или рассматриваемой предметной области. Исходя из этого, в работе изучается процесс построения формальной модели распознания значений текстовых единиц.

Как мы полагаем, в использовании современной компьютерной техники в Республике Казахстан значительная роль должна отводиться созданию лингвистических процессоров, основанных на применении знаний, заложенных на естественном казахском языке. Благодаря этому, современные компьютеры не просто будут осуществлять накопление и поиск информации на национальном языке, но и будут понимать смысл запроса, а также выдавать результаты в виде текста и речи на казахском языке. Создание таких лингвистических процессоров будет опираться на ограниченные языки определенных предметных областей (подъязыки). Выбирая наиболее частые и информационно насыщенные единицы языка, а также типовые контексты их употребления, мы можем построить такую модель языка (базовый язык), которая явится некоторым приближением к реальной системе подъязыка, порождающей тексты узкой тематики.

Особую ценность в республике в настоящее время представляет так называемая «деловая проза», написанная на казахском языке, а также армейские документы («военная деловая проза») и национальный военный лексикон, который только начинает зарождаться в армейской жизни.

Научные исследования в этом направлении, т.е. анализ и синтез текстов указанных документов, как мы представляем, являются первостепенной задачей ученых-тюркологов и специалистов по инженерной лингвистике.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Ахабаев А., Бектаев К.Б., Джубанов А. и др. Частотные словари и информационные характеристики языка//Тезисы докладов научно-теоретической конф. профес.-препод. состава Чимк. пед. инст., – Чимкент, 1967. – С. 107-108.

2. Джубанов А., Зубов А.В. Автоматизация некоторых лингвистических процессов // Вестник АН КазССР, – Алма-Ата, 1968, №9, – С. 31-36.

3. Джубанов А. К вопросу об автоматизации лингвистических исследований // Материалы первой научной конференции молодых ученых АН КазССР. – Алма-Ата: Наука, 1968. – С. 442-443.

4. Джубанов А. О некоторых путях вскрытия вероятностно-статистических закономерностей в современном казахском языке //

Материалы всесоюз. конф.: Статистическое и информационное изучение тюркских языков (СИИТЯ), – Алма-Ата, 1969. – с. 85-86.

5. Джубанов А., Ахабаев А., Мирзабеков С. О частотности формообразовательных аффиксов существительных казахского языка // Материалы всесоюз. конф.: СИИТЯ, - Алма-Ата, 1969. – С. 50-51.

6. Джубанов А., Шойбеков У. Статистическа аффиксов имен прилогательных // Материалы всесоюз. конф.: СИИТЯ, Алма-Ата, 1969. – С. 47-50.

7. Ахабаев А., Бектаев К.Б., Джубанов А., Мирзабеков С. Структурно-статистическая модель описания грамматических категорий тюркских текстов // Материалы Всесоюз. науч. конф. Часть I: Структурно-математические методы моделирования языка, Киев, 1970. – С. 13-14.

8. Бектаев К.Б., Джубанов А., Зубов А.В. Автоматическое построение частотных словарей // Вестник АН КазССР, Алма-Ата, 1970. – С. 48-53.

9. Джубанов А. Универсальная программа получения на ЭВМ «Минск-22» алфавитно-частотного списка сочетаемости графем // Материалы второй научной конференции молодых ученых АН КазССР, Алма-Ата, 1970. - С. 512-513.

10. Ахабаев А., Бектаев К.Б., Джубанов А. Исследование статистической структуры тюркского текста // Материалы всесоюз. конф.: Автоматическая переработка текста методами прикладной лингвистики. Кишинев, 1971. – С.17-18.

11. Джубанов А.К., Зубов А.В., Мирзабеков С. Информационное и статистическое изучение тюркских языков // Вестник АН КазССР, – Алма-Ата, 1971. – С. 82-84.

12. Жұбанов А., Мырзабеков С. Сөздік қорымыз қанша? // Білім және еңбек, Алматы: 1971, №8, – С. 14-15, 28-29.

13. Джубанов А.Х. Статистическое исследование казахского текста с применением ЭВМ (на материале романа М. Ауэзова «Абай жолы»): Автореф. дисс. канд., фил. наук. – Алма-Ата, 1973. – 37 С.

14. Бектаев К.Б., Джубанов А.Х. Вероятностные законы распределения классов слов казахского текста // – Трудах VII Всесоюзной школы-семинара: «Автоматическое распознавание слуховых образов», Алма-Ата, 1973.–С.139-145.

15. Джубанов А.Х. Становление статистического метода в казахском языкознании // Статистика казахского текста. – Алма-Ата: Наука, 1973. – С. 14-22.

16. Джубанов А.Х., Жубанов Е. Статистика лексико-морфологических форм прилагательных в романе М.Ауэзова «Абай жолы» (“Путь Абая”) // Статистика казахского текста. Алма-Ата: Наука, 1973. – С. 231-248.

17. Бектаев К.Б., Джубанов А.Х. Индексация и кодирование текстовой информации для ввода в ЭВМ // Статистика казахского текста. Алма-Ата: Наука, 1973. – С. 250-262.

18. Джубанов А.Х. Задача получения на ЭВМ частотных списков лингвистических единиц // Статистика казахского текста. Алма-Ата: Наука, 1973. – С. 263-298.

19. Бектаев К.Б., Джубанов А.Х. Вероятностно-статистическое моделирование тюркских текстов // – Статистика казахского текста. Алма-Ата: Наука, 1973. – С. 299-328.

20. Бектаев К.Б., Джубанов А.Х. О частотных списках графемно-фонемных сочетаний // Статистика казахского текста. Алма Ата: 1973. – С. 612-629.

21. Агманов Е., Джубанов А.Х. Распределения частот появления сочетаний знаков в Орхоно-Енисейской письменности. – В кн // Статистика казахского текста. Алма Ата: Наука, 1973. – С. 630-634.

22. Джубанов А.Х. Кибернетика и лингвистика // Материалы третьей научно-теоретической конференции молодых ученых АН КазССР. Часть II. Алма-Ата: Наука, 1974. – С. 245-247.

23. Бектаев К.Б., Пиотровский Р.Г., Джубанов А.Х. Информационно-статистическая типология тюркского текста // Советская тюркология и развитие тюркских языков в СССР // Тезисы докладов и сообщений Всесоюзной тюркологической конференции. Алма-Ата: Наука, 1976. – С. 305-307.

24. Құрышжанов А.Қ., Жұбанов А.Қ., Белботаев А.Б. Куманша-қазақша жиілік сөздік. Алматы: Ғылым, 1978. – 217 С.

25. Джубанов А.К. К вопросу о графемной статистике казахского текста // Вопросы казахской фонетики и фонологии. Алма-Ата: Наука, 1979. – С. 79-86.

26. Бектаев Қ.Б., Жұбанов А.Қ., Мырзабеков С., Белботаев А.Б. М.Әуезовтің «Абай жолы» романының жиілік сөздігі. Алматы: Ғылым, 1979. – С. 336.

27. Джубанов А.Х. О ходе подготовки к составлению словаря языка М.О.Ауэзова (функции ЭВМ). Материалы научного семинара: «Статистическая оптимизация преподования языков и инженерная лингвистика». Чимкент: 1980. – С. 193-195.

28. Джубанов А. Преобразование с помощью ЭВМ казахских текстов в фонетическую запись // Материалы семинара: «Статистическая оптимизация преподования языков и инженерная лингвистика». Чимкент: 1980. – С. 317-318.

29. Бектаев К.Б., Джубанов А.Х. и др. Использование ЭВМ в авторской лексиографии // Тезисы докладов Всесоюзной конференции: Переработке текста методами инженерной лингвистики. Минск: 1982. – С. 135-137.

30. Джубанов А.Х. Квантитативная структура казахского текста. Алма-Ата: Наука, 1987. – 147 С.

31. Джубанов А.Х., Ускомбаев С. А. О создании машинного фонда казахского языка // Материалы рабочего совещания: Машинные фонды языков народов СССР. Тбилиси: 1988. – С. 16-17.

32. Джубанов А.Х. О проблеме создания машинного фонда тюркских языков // Материалы рабочего совещания: Машинные фонды языков народов СССР. Таллин: 1988. – С. 4-5.

33. Джубанов А.Х., Кажибеков Е.З., Ускомбаев С.А. Создание словарной базы данных обще– и межтюркских моносиллабов: к проблеме разработки машинного фонда тюркских языков // Третья Всесоюз. конф. по созданию машинного фонда русского языка. Часть II. Москва: 1989. – С. 129-132.

34. Белботаев А.Б., Джубанов А.Х., Сиразитдинов З.А. О лингвистичсских исследованиях с применением вероятностно-статистических методов в тюркских языках // Статистика казахского текста. Выпуск II. Алматы: Ғылым, 1990. – С. 21-31.

35. Жұбанов А.Қ. Түркі тілдерінің машиналық қорын жасаудың алғышарттары // Статистика казахского текста. Выпуск II. Алматы: Ғылым, 1990. – С. 31-42.

36. Жұбанов А.Қ. Текст лингвистикасы жайында // Статистика казахского текста. Выпуск II. Алматы: Ғылым, 1990. – С. 118-125.

37. Ускомбаев С.А., Джубанов А.Х. Автоматический анализ морфологии казахской словоформы // Статистика казахского текста. Выпуск II. Алматы: Ғылым, 1990. – С. 148-156.

38. Джубанов А.Х., Ускомбаев С.А. Алгоритмизация задачи получения на ЭВМ фонетической транскрипции казахского текста// Тезисы докладов I Всесоюз. конф.: «Тюркская фонетика-90». Алма-Ата, 1990. – С. 35-36.

39. Джубанов А.Х. Использование ЭВМ для выделения опорных слов текста // Тезисы докладов Всесоюзной конференции: Актуальные проблемы компьютерной лингвистики // Тартуский универ. Тарту, 1990. – С. 49-50.

40. Джубанов А.Х. Автоматический отбор опорных слов текста в целях оптимизации обучения казахскому языку // Тезисы докладов Всесоюзной конференции: Использование ЭВМ в научной и учебной работе гуманитарных вузов. Минск, 1990. – С. 67-68.

41. Джубанов А.Х., Ускомбаев С.А. Алгоритмизация процедуры морфологического членения казахской словоформы // Тезисы докладов Всесоюзной конференции: Использование ЭВМ в научной и учебной работе гуманитарных вузов. Минск, 1990. – С. 87-88.

42. Джубанов А.Х., Зубов А.В. Формализация процедуры построения текстов естественного языка // Научные приборы и автоматизация научных исследований. Алма-Ата: Ғылым, 1992. – С. 190-196.

43. Бектаев Қ., Жұбанов А.Қ., Мырзабеков С., Белботаев А. М.О.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері. Алматы-Түркістан: 1995. – 346 С.

44. Жұбанов А.Қ., Мырзабеков С. М.Әуезов тілінің сөздіктерін жасау тәжірбиесі // Вестник КазГУ. Серия филологический., Алматы: 1997, №8. – С. 18-28.

45. Жұбанов А. Қ. Машиналық аударма // Қазақ тілі. Энциклопедия. IDK - TIPO. Қазақстан даму институты. Алматы, 1998. - Б. 268-269.

46. Жұбанов А. Қ. Математикалық тіл білімі // Қазақ тілі. Энциклопедия. IDK - TIPO. Қазақстан даму институты. Алматы, 1998. - Б. 267.

47. Жұбанов А.Қ. Проф. Қ.Жұбанов және қазақ тілін зерттеудің кейбір қолданбалы тәсілдері // Материалы республиканской научн. конф. «Жубановские чтения» - «Жұбанов тағылымы». Ақтобе, 1998. – С. 51-56.

48. Жұбанов А.Қ. Қ.Жұбанов және қазақ мәтін мазмұнын статистикалық тәсілмен анықтау // Материалы республ. научно-теоретической конф., посвящен. 100-летию К.Жубанова: «Құдайберген Жұбанов және қазақ тіл білімік. Алматы: Арыс. 1999. – С. 102-108.

49. Жұбанов А.Қ. Проф. Құдайберген Жұбановтың буын теориясы және оның компьютерлік лингвистикадағы орны // Материалы республ. научно-практической конф., посвящ. 100-летию К. Жубанова: Мемлекеттік тіл: терминология, іс қағаздары мен бұқаралық ақпарат құралдарының тілі. Астана: 1999. – С. 298-305.

50. Жубанов А.К. Основные процедуры формализации содержания казахского текста // Материалы межвузовской научно-практической конференции (3-4 февраля 2000) «Актуальные проблемы теории и практики перевода в контексте современности». Алматы, 2000. – С. 249-252.

51. Жубанов А.К. К проблеме семантической классификации знаменательных классов слов казахского языка // Известия МОиН РК НАН РК, Алматы, 2001, №2. – С. 13-30 .

52. Жубанов А. Статистическое исследование казахского текста на уровне абзацев // Тілтаным / Языкознание, Алматы, 2001, №2. – С. 35-41.

53. Жубанов А.К. Формальное описание словарной статьи автоматического словаря казахского языка (АСКЯ) // Тілтаным / Языкознание. Алматы, 2002, №2. – С. 26-40.

54. Аскар Кудайбергенулы Жубанов. Основные принципы формализации содержания казахского текста. Алматы, 2002. - 250 С.

55. Жұбанов А.Қ. Қазақ мәтінінің мазмұнын формализациялаудың негізгі принциптері //Материалы Республиканской научно-теоретической конференции «Жұбанов тағылымы-V». Актобе, 2002. – С. 40-42 (на рус.яз.).

КОМПЬЮТЕРЛІК ЛИНГВИСТИКА ЖӘНЕ ҚАЗАҚ ТІЛ БІЛІМІ

ҚР БжҒМ А.Байтұрсынұлы атындағы Тіл білімі институтының «Тілтаным» журналы. 2003. № 3. 28-36-бб.

«Компьютерлік лингвистака» термині (computational linguistics), негізінен, тіл қызметін модельдеуге қатысты пайдаланылатын компьютерлік бағдарламалар мен тілдік деректерді өңдеу, сол үшін қажетті компьютерлік технологияны жинақтау және компьютердің араласуына қатысты басқа да жұмыстармен байланысты қолданылып жүр. Кең мағынада алғанда, тіл білімі аясына қатысты теориялық және практикалық мәселелердің шешімін табуға

байланысты компьютердің араласу жағдайларының бәрі бірдей компьютерлік лингвистика саласына жатқызылады.

Адам баласының ақпараттық ортада өмір сүретіні белгілі. Ал, бұл ортада қабылдаушының жауапты ақпарат бойынша дұрыс шешімге келуіне жағдай жасайтын ақпараттық технологияның түрлері көп. Олар: автоматтандырылған ақпараттық жүйе, көпшілікке арналған коммуникация құралдарына жататын теле-радио байланыс жүйелері, компьютерлік аударма жүйесі, диалогтік жүйе және сұрақ-жауап, т.б. жүйелер. Мұндай жүйелердің саны мен сапасы күннен-күнге артуда. Ақпараттық қоғамдағы тілдік практика мен қазіргі кезеңдегі тілді зерттеу теориясының қолданылу аясы да өсіп келеді.

Тілді зерттеудегі компьютерлік лингвистика саласы қолданбалы тіл білімі атқаратын қызметтердің бір тармағы ғана деуге болады, себебі, «қолданбалы тіл білімі» ұғымы көптармақты бола келе, күнделікті өмірде дәстүрлі және жаңаша мағынада кездесіп жүр. Мәселен, бұл ұғым дәстүрлі түсінік бойынша, тіл қолданым мәселелерінің практикалық әдістерін жасауды, яғни жазу мәселесін дамытуды, ауызекі сөйлеудің транскрипциялық жүйесін құруды, шетел сөздерінің транскрипция жүйесін жасауды, стенография тәсілдерін дамытуды, арнайы лингвистикалық сөздіктер шығаруды, тарихи ескерткіштердегі таңба-жазуларды айқындауды, ғылыми-техникалық терминологияны қалыптастыруды, т.б. іс-шараларды қамтиды.

Қолданбалы тіл білімінің жоғарыда сөз болған дәстүрлі саласы қазақ тілінде ойдағыдай дамып, түркі тілдері ішінде көш бастап келеді деуге болады. Ал, оның компьютерге байланысты жаңа саласы – «компьютерлік қазақ тіл білімі» әлде де кешеуілдеп дамуда.

Қолданбалы тіл білімінің «статастикалық лингвистика» (квантитативті лингвистика) тарауы бойынша ғылыми ізденістер проф. Қ. Бектаевтың атымен және оның ғылыми мектебімен тығыз байланысты. Қазақ тілінің әр түрлі стильдерінің мәтіндері бойынша, алғашында ЭЕМ-ның, кейіннен компьютерлердің көмегімен алынған жиілік сөздіктер (әліпбилі жиілік, жиілік, кері әліпбилі) тілдің құрылымдық жағының статистикасы жайында көптеген мәліметтер алуға мүмкіндік туғызды. Осылайша, қазақ тіліндегі бай статастикалық мәліметтер негізінде бірнеше кандидаттық және докторлық диссертациялардың қорғалғаны көпшілікке мәлім. Бүгінгі таңда А.Байтұрсынұлы атындағы Тіл білімі институтының «компьютерлік лингвистика» бағыты бойынша қол жеткен ғылыми нәтижелері айтарлықтай деуге болады. Осылардын тек негізгілері деп мынадай ғылыми жұмыстарды атауға болады:

– көркем әдебиет, көсемсөз, ғылыми-көпшілік әдебиет пен оқулықтар мәтіндерінің, алғашында ЭЕМ-ның, кейіннен компьютердің жадына көптеп енгізіліп, олар бойынша арнайы компьютерлік бағдарламалар құрастырылуының негізінде әр түрлі жиілік сөздіктердің алынуы;

– М.Әуезовтің 20 томдық шығармалар жинағының компьютер жадына енгізілуі және соның негізінде жазушы тілінің жиілік сөздіктерінің жарық көруі;

– осы тәрізді жиілік сөздіктердің негізінде қазақ сөздерінің әр стильде және әр көлемді мәтіндерде қайталануының қосынды жиілігінің, сол мәтіндерді қамту статистикасының анықталуы;

– сөзтұлғалардың құрылымдық статистикасы мен мәтіндегі пайыздық салмақтарының әр сөз таптарына қатысты айқындалуы;

– қазақ тіліндегі негізгі сөз таптарының мәтін бойында кездесулерінің математикалық статистика мен ықтималдық теория заңдылықтарына бағыну-бағынбау жағдайларының зерттелуі;

– қазақ тілінің «ТІЛ – ҚАЗЫНА» атты автоматтандырылған компьютерлік картотекалық қорының іске қосылуы;

– қазақ тіліндегі мәтін мазмұнын оның тұрпатына қарай ашудың (тұрпаттаудың, формалдаудың) негізгі принциптерінің айқындалуы;

– Тіл білімі институтында жоспарлы тақырыптар бойынша түзіліп жатқан түсіндірме, екі тілдік, бір тілдік сөздіктер мәтіндерінің компьютер жадында «сөздіктер қорын» құрап, тұрақты сақталуы.

Осы аталған іс-шаралардың қатарын, компьютердің араласуымен орындалған жұмыстардың және компьютерлік тіл білімі саласына қатысты басқа да жұмыстардың нәтижелерімен ұштастыруға болар еді. Бірақ, ол мақаламыздың мақсатына кірмейді.

Мақалада негізінен түркі тілдері үшін тіл зерттеудің жаңа саласы деп есептелгенімен, үнді-еуропа тілдерінде үйреншікті болып, күнделікті зерттеу мәселелеріне айналған компьютерлік лингвистиканың кейбір жәйттеріне тіл зерттеудегі жаңаша көзқарасқа қызығушы қауымға бағыштап, арнайы сөз етуді жөн көрдік.

50-жылдардан бастап «қолданбалы лингвистика» термині жаңа мағыналарға ие бола бастады. Ол күнделікті өмірімізге компьютерлік технологияның және автоматты басқару жүйесінің араласуы арқылы мәтінді автоматты түрде өңдеуге қатысты көптеген қолданбалы бағыттағы тілдік бірліктерді талдау (анализ) мен жинақтау (синтез) негізінде тілдің танылуы мен талдану заңдылықтарын белгілеу қажеттігінен туындайды. Орыс тілінде жарық көрген әдебиеттерде «қолданбалы лингвистика», «есептеуіш лингвистика», «инженерлік лингвистика», «автоматты лингвистика», «математикалық лингвистика», «статистикалық лингвистика» терминдері бірінің орнына бірі қолданыла бастады. Әрине, бұл аталған терминдерді ортақтастыру дұрыс шешім деуге болмайды, себебі олардың қолданбалы тіл саласындағы мақсаты мен міндетінде және әдіс-тәсілдерінде өзіндік ерекшеліктері баршылық. Тілдің қолданбалы шеңберінің кеңеюін оның әр түрлі ғылымдар салаларымен түйісу мүмкіншілігінен деп түсінген жөн [1, 23]. Тіл қызметінің пәнаралық сабақтастық қасиеті тілдік теорияға тек қана оң әсерін тигізумен қатар, қазіргі тіл білімінің тұжырымдамалық бағытын (концептуальдық аппаратын) жаңартуға да мүмкіндік туғызды.

«Қолданбалы лингвистика» саласының кең мағыналылығын – тілдің құрылымы мен қызметіне қатысты ғылыми мағлұматтарды тілге қатыссыз пәндер саласында, адам баласының практикалық мұқтаждығынан және мұндай қызметтің теориялық негіздемесі ретінде қолданылуынан туындайтынынан көруге болар еді.

Қолданбалы лингвистиканы функционалдық көзқарас тұрғысынан қарастырсақ, оны тіл қызметінің ең қолайлы (оптимальды) әдіс-тәсілдерін тұтынушыға ұсынатын ғылыми пән деп те түсінуге болады. Тілдің коммуникативті функциясының қолайлы жақтарын ұсынатын пәндер – аударма теориясы, машиналық (компьютерлік) аударма, ана тілі мен шет тілдерін оқытудағы «ақпараттық іздестіру» жүйесінің теориясы мен практикасы, ақпараттық жасанды тіл жүйесін жасаудағы таңбалау теориясы.

Тілдің коммуникативті функциясының бір көрінісі ретіндегі әлеуметтік қызметі әлеуметтік тіл білімінде, тілдік жоспарлау мен тілдік саясатта, орфография мен орфоэпияда, әсерлік теория мен саяси лингвистакада көрініс табатыны белгілі. Сол сияқты тілдің когнитивті қызметінің ең қолайлы жақтары компьютерлік тіл білімінде, психолингвистика мен афазиологияда және квантитативтік лингвистика салаларында көрініс тауып жүр [2, 8].

Компьютерлік лингвистика информатиканың тілдік негіздемесін қарастырумен қатар «тіл» мен «ойлау» арасындағы байланысты және одан туындайтын тікелей болмысты компьютер арқылы модельдеуге қатысты мәселелермен шұғылданады. Компьютерлік лингвистиканы тілдік бірліктерге қатысты проблемалар да қызықтырады. Ондай бірліктер сөзден кіші, үлкен және оған тең де болуы мүмкін. Бұл жердегі сөзден үлкен деп отырғандарымыз – сөз тіркестері, сөйлемдер (айтылымдар) және толық мәтіндер. Аталып отырған салада тілді зерттеу кезінде қолданылатын амал-тәсілдерді модельдеу жолдары да қарастырылады. Мәселен, мәтін мазмұнын ашу жолдарын немесе аудармашының бір тілдегі мәтінді екінші тілге аудару жолдарын модельдеу мәселелері бұған мысал бола алады.

Сөйлеу мен ойлау процестерінің арасындағы байланыстардың аса күрделі философиялық мәселе екені мәлім. Біз олардың түрлерін көптеген айтылымдардағы, сөйлемдер мен мәтіндердегі тілдік мәлметтерді тұрпатына қарай (формалды) сипаттауда ұтымды пайдалануға болады демекпіз.

Адам баласының «ойлау» қабілетін модельдеу жағдаятында ғалымдардың назары көбінесе лингвистика мен логика арасындағы байланысқа ауып, тұжырымдау процесі ойлаудың анықтаушы және шешуші сипаты деп саналады. Қазіргі тіл білімінде тіл мәселесіне байланысты кейбір жәйттерді логика саласының мамандары қарастырып, ал кейбір логикалық тұрпаттар (формалар) сырын ашу жағы тілшілерге жүктелуде. Бұл түсінікті де. Себебі, грамматикалық (семантикалық) және логикалық тұрпаттар әр көзқараспен қаралатын бір ғана нысанның түрлері болып келеді. Табиғи тіл адам баласын дағдылы ақиқат шындығынан уақытша ажыратып, тілдік тұлғалардың семантикалық мағынасымен анықталатын «ойлау» тұрпаттарымен (формаларымен) әрекет жасайтын жаңа деңгейге ауысуына жағдай жасайды.

Бұл айтылғандардың барлығы да компьютерлік лингвистика үшін аса маңызды мәселелер. Себебі, адам баласының ойлау қабілетін тұрпатына қарай (формалды түрде) әр түрлі дәлдікпен қайталаудың мүмкіншілігі туады. Ал, бұл жағдай математикалық модельдеу мен машиналық логиканың көмегімен іске асатын «ойлау» қабілеті бар «жасанды интеллект» жасап шығаруға мүмкіндік туғызады. Бұл жердегі «жасанды интеллект» атты жүйені – іс жүзінде «ойлау» қабілеті бар және «шығармашылыққа» қатысты мәселелердің шешімін таба алатын компьютерлік бағдарлама (программа) деп түсінген жөн.

«Жасанды интеллектіні» іске қосу үшін ең алдымен адам баласының белгілі бір жағдайда шешім қабылдаудағы немесе жауапты мәселеге шешім іздеу кезіндегі ойлау процесінің сырын зерттеп тануы қажет. Әсіресе, адамның ойлау процесіндегі көптеген сатылардың ішінен ең маңыздыларын бөліп алуды мақсат еткен жөн. Бұл жолда күрделі ойлау қызметінің құрылымы жағына мән беріле отырып, шығармашылықтың құпиясын қарапайым сатылы бөліктерге бөлшектеп, компьютерлік бағдарламаның алгоритмін құрудың, яғни болашақ «ойлау» қабілеті бар «жасанды интеллект» жасап шығарудың алғашқы кірпішін қалау жолдары зерттеледі.

Ми қызметін дәлме-дәл анықтау мүмкін бола бермейді, себебі, адамға тән ақыл-ой – өте күрделі әлеуметтік-биологиялық функция, сондықтан да оның құпиясын ашу көптеген зерттеулерді қажет ететіні айқын.

«Жасанды интеллектіге» қойылатын негізгі талап ми қызметінің моделін «ойлау құралы» деп ұйғарудан туындайды. Тіл қызметіне қатысты модельдеудің негізгілері болып есептелетін – аударма мәселесі. Бұл тілдік құрылымның морфологиясынан синтаксиске дейінгі барлық деңгейін қамтиды. Сондықтан да компьютерлік аударма тіл қызметінің аса күрделі функцияларының бірі болғандықтан, оны модельдеу де оңайға түспейді. Қазіргі кезде көптеген еуропа тілдерінен орыс тіліне және керісінше компьютерлік аударма жасау мәселесі біршама шешілді деуге де болады. Ал, осы сияқты аудармаларды түркі тілдері бойынша, оның ішінде қазақ тіліне қатысты іске асыру – егеменді еліміздің кезек күттірмейтін мәселелерінің бірі болып отыр. Қазақ тілінің компьютерлік саласы жаңа ғана дамып келеді және оның ойдағыдай дәрежеге жетуіне барлық мүмкіндік бар. Себебі компьютерлік лингвистика саласы жеке ғылыми бағыт ретінде 60-жылдарда-ақ қалыптасты. Бұл сала бойынша жарық көрген ғылыми жұмыстар баршылық. Мәселен, АҚШ-та әр тоқсан сайын «Компьютерлік лингвистика» атты ғылыми жорнал жарық көреді. Осымен байланысты компьютерлік лингвистика мәселесі бойынша айналысатын аймақтық құрылымдағы әлемдік «Қауымдастықты» да айта кетуге болады. Осымен бірге әрбір екі жыл сайын осы сала бойынша өткізілетін «КОЛИНГ» атты халықаралық конференцияның жұмысы да айтарлықтай жемісті деуге болады [2, 14].

«Компьютерлік лингвистиканың» айналысатын мәселелерінің ең маңыздыларының бірі – табиғи тілдің мәтін морфологиясын автоматты түрде талдау. Егер, дәстүрлі түсінік бойынша «сөз морфологиясы» деген ұғым сөздің тұлғалық құрылымын, яғни сөз түбірін және оған жалғанатын

қосымшалардың түрін анықтайтын болса, компьютерлік тіл біліміндегі «морфология» терминінен басқаша ұғым туындайды. Дәлірек айтсақ, ол сөздің мәтін бойындағы сыртқы пішініне (тұрпатына) қарай, тілдік құрылымның әр қилы деңгейлері бойынша тілдік мәліметтер алу мүмкіндігін білдіреді. Алғашында, «морфологиялық талдау» ұғымы айтылған мағынада машиналық аударма саласында пайда болды [3].

«Морфологиялық талдаудың» жанаша түсінігі бойынша алынатын ақпарат біздің дәстүрлі ұғымымыздағы морфологияға қатысы жоқ көптеген амалдардан тұруы мүмкін. Осының салдарынан компьютерлік тіл біліміндегі «морфологиялық талдау» ұғымы – амалдар ұғымы (операционное понятие), яғни сөздің сыртқы пішіні бойынша тілдік ақпараттарды танып-білу үшін жүргізілетін іс-әрекет деп ұққан жөн сияқты. Басқаша айтсақ, егерде дәстүрлі тіл білімі ұғымы бойынша «морфологиялық талдау» кезінде «нені талдаймыз?» деп сұрақ қоятын болсақ, компьютерлік тіл білімінде «қалай талдаймыз?» деген сұрақ қойылады, яғни ол сөз тұлғасы бойынша қажетті тілдік ақпараттарды қай жолмен алуға болатындығымен айналысады.

Машиналық (компьютерлік) аударманың алғашқы тәжірибелерінде мұндай амалдардың саны көптеп кездесетін, ал бүгінгі таңда орыс және еуропа тілдері бойынша бұл мәселе біршама шешілді деуге болады [4].

Қазақ тілі мен басқа да түркі тілдері бойынша мұндай ізденістер айтарлықтай қолға алынбай жүр.

Аталған түсініктегі «морфологиялық талдау» бірнеше бағытта құрастырылуда. Одардың біріншісі – сөзтұлғаны оның негізі мен оған жалғануы мүмкін болатын қосымшаларға ажырататын классикалық талдау негізінде модельдеу.

Екінші бағыт сөзтұлғалардың соңғы әріптерінің тіркесімдік заңдылықтарының жиілік сөздіктердегі статистикалық мәліметтеріне сүйенеді.

Үшінші бағыт – соңғы кездердегі ізденістердің нәтижесі. Бұл бағыт бойынша теңдеулердің ашық жүйелерінің пішіні ретінде, морфологияның әмбебап математикалық моделі жасалады. Модельдің есептегіш мүмкіндігі негізінде сөзтұлғаларды нормалау және қажетті грамматикалық ақпарат алумен қатар, сөзтұлғаларды жинақтау (синтездеу) да іске асады [5, 44].

Сөзтұлғалардың әріптік құрамының өзгеруін анықтайтын сөздерді өзара топтауды (сыныптарға бөлуді) негіз етіп, автоматтандырылған морфологиялық талдаудың алгоритмі құрастырылады. Мұндай топтау «морфологиялық топтау» деп аталып жүр [6]. Сөзтұлғалардың жазылуындағы әріптік өзгеріске ұшырау флективті және агглютинативті тілдерде біркелкі емес. Мәселен, орыс тіліндегі: «сижу – сидишь» сөзтұлғаларының жазылуындағы сөз негізіндегі әріптер өзгеріске ұшыраса, сол сөздердің қазақша баламасы – «отырмын – отырсың» сөздерінің жазылуында қосымшадағы әріптер өзгерген.

АМПАР атты машиналық аударма жүйесінде сөздердің синтаксистік қызметі мен септік және т.б. жалғаулардың негізінде сөздердің

морфологиялық кластары екі түрге бөлініп қарастырылды: 1) негіздерінде өзгеріс болатын сөздер тобы, 2) сөздердің флективті тобы.

Соңғы топ, белгілі сөздерге ғана тән қасиеттегі белгілер жүйесі арқылы немесе осындай қасиеттерді бойына сақтаған дерек сөздер арқылы сипатталады.

Енді морфологиялық талдаудың түрлеріне қысқаша тоқталайық. Олар төмендегідей:

- негіз сөздердің сөздігі арқылы морфологиялық талдау жүргізу;
- сөзтұлғалар сөздігі көмегімен морфологиялық талдауды іске асыру;
- логикалық көбейту әдісін қолданып, морфологиялық талдау жасау;
- морфологиялық талдауды сөздіктердің көмегінсіз (сөздіксіз) арнайы кестелер арқылы іске асыру.

Еуропа тілдерін зерттеуде көп тараған морфологиялық талдау түрі – негіз сөздердің сөздігі мен кейбір көмекші кестелер жүйесін пайдалану. Аталған сөздіктерде ішкі флексиясыз, жәй және күрделі сөздердің негіздері мен олардың негіз тұлғалары (формалары) толық беріліп отырады. Сөздіктегі әрбір негіз сөзге морфологиялық кластардың екі түрін ажырататын шартты белгі – «код» қойылады. Ал, омонимдік негіздерге шартты белгілердің (кодтардың) тіркесімдік түрлері беріледі.

Морфологиялық талдаудың екінші түрі, яғни сөзтұлғалар сөздігі арқылы жүргізілетін талдау да зерттеушілер тәжірибесінде көп қолданыс тапқан, кең тараған талдау түрі болып саналады. Олай дейтініміз, морфологиялық талдаудың компьютерлік алгоритмінде морфемаларға бөлшектеу мен оларды сөздік бойынан іздестіру әрекеттері негіз сөздер сөздігіне қарағанда жеңіл іске асады. Бірақ талдаудың бұл түрінің өзіне тән осал жерлері де жоқ емес. Мысалы, егер іздестіретін сөз сөзтұлға сөздігінде кездеспеген жағдайда, біз оның грамматикалық ақпараты жайлы да ештеңе біле алмаймыз. Сондықтан сөзтұлға сөздігі көмегімен талдау жүргізу жүйесінде аффикстер мен түбір сөздер тізімдері берілуі қажет. Осымен бірге бірнеше сөзтұлғаларды бір лексикалық бірлікке сәйкестендіруге қажетті, оларға тән қасиеттерге ие, белгілер (атрибуттары) толық берілгені талап етіледі. Бұл айтылғандардан туындайтын қорытынды – морфологиялық талдаудың негіз сөз сөздіктері арқылы жүргізілгені ұтымды.

Морфологиялық талдаудың логикалық көбейту әдісі зерттеу аясынан айырықша орын алады. Бұл әдісте «сөздік функциясы» деген ұғымға ерекше көңіл бөлген жөн. Әрбір «сөздік функциясына» сөзтұлға функциясы және оған тән ақпарат сәйкестендіріледі. Осының негізінде әр сөзтұлғаның өзіне ғана тән ақпараттық деректер арқылы функция мәнін арнайы кестемен беруге мүмкіндік туады. Бірақ, аталған функцияның мәнін басқаша, төменде көрсетілетіндей тиімді амалдар арқылы беруге де мүмкіндік бар:

1) сөзтұлға – әріптер тізбегі ретінде морфемдік сегменттерге бөліктенеді;

2) сөзтұлға – морфемдік сегменттер тізбегі ретінде, басқа морфемдік элементтердің реттелмеген жиынымен ауыстырылады;

3) сөзтұлғаға, морфемдік жиын ретінде, белгілі-бір ақпарат сәйкестендіріледі;

4) бұл ақпарат сөзтұлға жайындағы қажетті деген ең соңғы ақпарат түрінде қайтадан өзгертіліп беріледі [5, 46].

Әрбір морфемаға құрылымында сондай морфемасы бар сөзтұлғалардың жиынтық ақпаратын сәйкес қоюға болады. Логикалық түсінік бойынша мұндай ақпараттар жиынтығы дизъюнкциямен, яғни сөзтұлғаның құрылымдық белгілерінің бір-біріне қарсы қойылуымен сәйкес келеді. Сөзтұлға жайлы ақпарат осы сөзтұлғаның табиғатына тән қиылысудан немесе логикалық конъюнкция сияқты морфемалық ақпараттардан тұратынына көз жеткізуге болады.

Морфологиялық талдаудың логикалық көбейту әдісі, көбінде, флективті тілдерде қолданылады және негіз сөздер сөздігінің болуын талап етеді.

Қарастырып отырған автоматтандырылған морфологиялық талдаудың соңғы түрі – сөздіксіз талдау немесе «тәуелсіз» талдау. Ол аффикстер кестелері көмегімен және арнайы түзілген грамматикалық мағынасы жоқ сөздердің тізімі негізінде іске асады. Морфологиялық талдаудың бұл түрі зерттеу тәжірибесінде өте сирек қолданыс тауып жүр.

Енді автоматтандырылған морфологиялық талдаудың қазіргі күйі қандай деген сұрақтың жауабына қысқаша тоқталайық.

Қазіргі кезде кез келген мәтіннен компьютер арқылы сапалы ақпарат алу жүйесіне қойылатын талап заман сұранысына қарай өсті. Енді компьютер көмегімен морфологиялық талдау жасау алгоритмін іске қосу мәселесіне, негізгілерін алғанда, мынадай талап қойылуда:

- кез келген тақырыптағы мәтіннің 98-99% қамтитын көптақырыпты (политематический) «аса күшті» сөздік морфологиялық талдау жүйесінің негізін құруы тиісті;

- автоматты түрдегі талдау алгоритмі кез келген сөзтүрленістерін ескеретіндей мүмкіндікке ие болуы қажет. Осы жағдай ескерілсе ғана сөзтұлғаны жан-жақты тану мүмкіндігі артады және сан жағынан алғанда ондай мүмкіндіктің нәтижесі сөздікте қамтылған лексикалық бірліктен бірнеше есе артық болуы мүмкін;

- автоматтандырылған морфологиялық талдау жүйесінде «жаңа» сөздер, сөздікте қамтылған басқа сөздермен бірдей дәрежеде қарастырылуы қажет және оларды дұрыс танудың ықтималдығы 90-95% кем болмауы қажет;

- мәтінді компьютер арқылы өңдеу жылдамдығы тұтынушыны қанағаттандыратындай болуы қажет және оған енгізілетін мәтін көлеміне шек қойылмауы керек;

- компьютер арқылы жүргізілетін морфологиялық талдау жүйесінде қажеттікке қарай өзгерістер жүргізу (сөздікті толықтыру, өңдеу және т.б.) мүмкіндігі алдын-ала ескерілуі қажет, яғни бұл аталған жүйені «үйретуге» барынша мүмкіндік тууы керек [5, 48-49].

Бүгінгі таңда, орыс тіліне және басқа да еуропа тілдеріне қатысты морфологиялық талдау жүйелеріндегі кейбір сөздіктер мен мәтіндердің көлемдері жайлы мынандай мәліметтерді айтуға болар еді: негіз сөздерден тұратын политематикалық сөздіктің көлем 100 мың лексикалық бірліктен кем емес; сөздік, көлемі 30 млн. сөзқолданыстан асатын политематикалық мәтін негізінде жасалғандықтан, оның кез келген тақырыптағы ғылыми-техникалық мәтінді қамту дәрежесі өте жоғары; көлемі 3 млн. сөзқолданыстан тұратын мәтіннің негізінде түзілген сөзтұлғалар сөздігі 46 мың, лексикалық бірлікті құрайды. Сөз болып отырған екі сөздікте де сөздердің құрылымы мен синтаксистік қасиеті жөнінде толық түрдегі грамматикалық ақпарат берілген [7].

Автоматтандырылған морфологиялық талдау сөздіктердің айрықша (жана, арнайы) түрлерін өмірге келтірді. Оларда тілдік бірліктердің тұлғасына, туындау ережелеріне, сөзтудырғыш, сөзтүрлендіргіш топтарына қарай реттеу жақтары қарастырылды. Осындай типтегі сөздікке тілдік бірліктің соңғы жағынан бастап әліпбиге түсірілген кері әліпбилік (обратно-алфавитный) сөздік деп аталынатын сөздікті жатқызуға болады. Морфологиялық талдаудың талабына сай түзілген бұл сөздікте сөзтудырғыштық, сөзтүрлендіргіштік типтегі сөздер және соңғы қосымшалары бірдей болып келетін күрделі сөздер өзара жіктелген ретте көрініс табады. Кері әлілбилі сөздік бойынша сөздердің морфологиялық құрылымына қатысты көптеген мәселелерді шешуге болады. Мәселен, қосымшалары бірдей болып келетін сөздердің топталып берілуі олардың грамматикалық сипатын және қосымша мен қандай сөзтүрлендіргіш типке жататындығы арасындағы қатынастарға қатысты деректерді анықтауға мүмкіндік туады. Грамматикалық форманттардың синонимдік, омонимдік қатарларын жеңіл ажыратуға және олардың тіркесімдік, сандық мөлшерінің сипаттарын анықтауға мүмкіндіктер туындайды. Кері әліпбилі сөздік бойынша, аналитикалық тілдерге қарағанда флектавті және агглютинативті тілдер жүйелерінен морфологиялық форманттарға байланысты ақпараттарды көптеп алуға болады.

Осы уақытқа дейінгі Қазақстандық «Статистика тобының» қазақ тілін ЭЕМ-ның және компьютердің көмегімен лингво-статистикалық зерттеу барысындағы қол жеткізген көптеген ғылыми нәтижелерін, яғни әр мақсатпен түзілген жиілік сөздіктер мен олар арқылы алынған құнды деректер мен тұжырымдарды әрі қарайғы іс-тәжірибемізде дұрыс пайдалана білуіміз керек. Ең алдымен, сол ғылыми жетістігімізді үнді-еуропа тілдеріндегі «компьютерлік лингвистика» саласы бойынша орындалып жатқан ғылыми-зерттеу жұмыстарының деңгейіне дейін көтеру жолында жұмыс жасау қажет.

Осы мақалада сөз болған компьютердің араласуымен ғана жүзеге асатын морфологиялық талдаулар мен басқа да зерттеу түрлерін мемлекеттік тілімізге пайдалана отырып, ана тіліміздің компьютерлік лингвистика саласын дамытуды және соның нәтижесінде қол жеткізген жетістігімізді күнделікті өмір қажетіне пайдалануды заман талабы деп түсінеміз.

Әдебиет

1. *Городецкий Б.Ю.* Актуальные проблемы прикладной лингвистики // Новое в зарубежной лингвистике. Вып. XII. 1983.
2. *Баранов А.Н.* Введение в прикладную лингвистику. Эдиториал УРСС, М., 2001.
3. *Василевский А.Л., Марчук Ю.Н.* Вычислительная лингвистика. Учебное пособие для студентов отделения прикладной лингвистики. М.,: МГПИИЯ им. М.Тореза, 1970.
4. Лингвистические вопросы алгоритмической обработки сообщения. М., 1983.
5. *Марчук Ю.Н.* Основы компьютерной лингвистики, Учебное пособие. М. 2000.
6. *Белоногов Г.Г., Новоселов А.П.* Автоматизация процессов накопления, поиска и обобщения информации. М., 1979.
7. *Зеленков Ю.Г.* Морфологический анализ в системах автоматической обработки научно-технической информации. Канд. дис. М.,: ВИНИТИ, 1988. С.145.

«АДАМ-КОМПЬЮТЕР-АДАМ» ЖҮЙЕСІНДЕГІ ТАБИҒИ ТІЛ ҚЫЗМЕТІНІҢ ЕРЕКШЕЛІКТЕРІ

ҚР ҰҒА-ның «Хабарлары». Тіл, әдебиет сериясы.
2003. № 4. 3-11-бб.

Қоғамның даму сатысы күрделене түскен сайын, ондағы ақпарат құралы қызметін атқаратын – «тілдің» дәрежесі де айтарлықтай көтеріледі.

Кезінде ауыл шаруашылығындағы жер жырту ісіне соқаны пайдаланған дәуір мен қазіргі компьютер және интернет заманындағы өндірістік қарым-қатынастар салыстыруға келмейтіндей болып өзгерістерге ұшырады. Қоғамның дамуы ақпарат ағымының үздіксіздігі мен сапалылығына өз әсерін тигізбей қоймады. Әсіресе, қысқа және ұзақ мерзімді ақпарат түрлері айтарлықтай өзгеріске ұшырады. Осының әсерінен, біріншіден, тілдік құрылымның негізгі қаңқасы ретіндегі грамматиканың, түбір сөз бен өнімді сөзжасам моделі жүйесінің тұрақтылығы, екіншіден, өндірістік қажеттілік пен қоғам өмірінің қажеттігіне сай сол қаңқаны толықтыратын лексикалық бірліктер мен олардың тіркесімдерінің сан жағынан ұлғайуымен бірге, оларда тұрақсыздық қасиеттер де байқалуда [1],

Адам өміріне электрондық техниканың келуі бұлшық етке қатысты көптеген жұмыстарды жеңілдетумен қатар, ми қызметінің де іс-әрекеттеріне араласа бастағаны бәрімізге мәлім. Осыған байланысты адам миының қызметіне бәсеке тудыратын неше түрлі «ақылды» автоматтар мен «ойлау» қабілеті бар электронды машиналар жайлы жәйттер соңғы кезде көптеген басылымдарда жарық көруде. Кейбір авторлар, тіпті осыдан 200 жыл бұрын

тоқу станогының тоқымашыны ығыстырғаны сияқты, электронды машиналар адамзатты ми қызметінің өрісінен де ысырып тастап, оған тек тұтынушының ғана функциясы қалдырылуы мүмкін деп те күмән туғызады. Әрине, мұндай қисынсыз болжам адам баласының ми қызметінің мәнін дұрыс түсінбей, оны түгелімен бөлшектелетін (алгоритмделінетін) процеске жатқызып, компьютер арқылы қайта өндіруге болады-мыс деп ұйғарудан туындайды. Шындығында, адамның ойлау қабілеті негізінен дедуктивті-логикалық емес, иңдуктивті-ықтималдық болғандықтан, толығымен бөлшектелуге жатпайтын процесс деп ұйғару керек. Бұл айтылғандардан адамның ойлау қабілетін түгелімен формалдап, компьютер жадында алгоритмдеу амалымен қайта өндірудің, яғни қайтадан қалпына келтірудің мүмкін еместігін ескертпекпіз.

Соңғы кезде ақпараттардың өсу көлемі жайлы жаңа көзқарас байқалады. Ол – қысқа мерзімді ақпараттың көлемі арифметикалық прогрессиямен, ал ұзақ мерзімді ақпарат – геометриялық прогрессиямен ұлғаяды деген көзқарастарды қолдау. Сонымен қатар, ақпараттық процестерді компьютерлендіру (автоматтандыру) ғылым саласынан гөрі, өндірісте көбірек және тезірек іске асады деп есептеу тенденциялары.

Табиғи тілге байланысты төмендегідей екі түрлі теориялық проблемаларды сөз етуге болады.

Олардың біріншісі – тілдің компьютер жадындағы қызметіне қатысты туындайтын сұрақ.

Табиғи тіл (ауызша не жазбаша) адам баласынан логикалық роботқа (компьютерге) түгелдей ауысу кезінде қандай сапалық өзгеріске ұшырауы мүмкін, әлде «тіл» өзгеріссіз алғашқы қалпында қала ма?

Екінші сұрақ – табиғи тілді зерттеу барысында формалдау процесінің шегі бола ма? Басқаша айтқанда, «тілді» мазмұнға қатыссыз бөлек алып суреттеу мүмкін бе? Егер олай қарастыруға болмаса, ондағы тілдік бірліктердің берілуі мен олардын мағынасының ара қатынасын анықтау болып табылады ма? - деген сұрақ туындайды.

Бірінші сұрақтың жауабы: «тіл» ұғымын әр түрлі тұрғыда қарастыруға байланысты түсіндіруге болады. Мысалы, «тілді» тек семиотика тұрғысынан анықтасақ, онда тіл – символдарға қатысты кодтар жүйесі деп есептелінеді. Бұл түсінік бойынша адам да, логикалық робот та кибернетикалық жүйе деп ұйғарылады. Осының негізінде «тілдің» адамнан роботқа немесе керісінше – роботтан адамға ауысу процесінде өзгеріс болмайды деп есептеу керек.

Енді «тілді» мазмұнды лингвистикалық тұрғыдан қарастырсақ, онда болмыста кездесетін «тілдің» екі түрі бар деп, яғни олардың жеке (дара) және ұлттық тіл түрлері болады деп ұсынған жөн. Тілдің кейбір қолданбалы есептерін шешуде тілге семантикалық көзқарас ыңғайлы болғанымен, ол жеткілікті емес. Себебі табиғи тілдің екі жақтылық табиғаты оны таңбалар жүйесінің теориялық аясында толық баяндауға мүмкшшілік бермейді. Екі жақтылық табиғат, біріншіден, жекелік және ұлттық жүйе түрінде байқалса, екіншіден, қарым-қатынастық (коммуникативтік) құрал ретінде көрініс табады.

Егер логакалық роботты (компьютерді) аса бір зор жазба кітапшасымен салыстыратын болсақ, оның ұқсастығымен қатар, айырмашылығын да аңғаруға болады. Ұқсастығы – екеуінде де табиғи тілдің жазба көрінісі болуында.

Компьютер мен адам арасындағы коммуникативтік (қарым-қатынастық) жүйе арқылы мәтіндік ақпараттың адамнан компьютерге ауысу процесінде табиғи тіліміз, оның көрінісі ретіндегі жазбамен алмасады, ал керісінше, ақпарат компьютерден адам назарына ауысқанда, жазба мәтінінің компьютердегі керінісі, қайтадан, таңбалардың психикалық кескін үйлесімінің бұрынғы қалпына келтіріледі, яғни ішкі тіл (внутренняя речь) қалпына ауысады.

Сонымен, кез келген мәтін бойынша компьютермен жұмыс істеу – тілге психикапық әсер ету емес, ол мәтіннің формалды шартты белгілерінің оның өз таңбаларымен әрекет етуі деп түсінген жөн сияқты. Бұл әрекеттер, әрине, адамның ойлау қабілетіндегі процестерге ұксай бермейді.

Табиғи тілдің құрылымын формалдаудың шегін анықтау қажеттігі – тілдің берілу пішіні (план выражения) мен мазмұндық көрінісі (план содержания) арасындағы қатынастарын арнайы зертеуді қажет етеді. Тілдік фактілердің формалдану байланыстарын зерттей келе, біз жазба тілдің берілу көрінісін баяндауда олардың формалдану заңдылығына бағынатынын байқаймыз. Ал, тілдік элементтердің берілу көрінісі мен мазмұндық берілуінің арасындағы байланысты зерттеу барысында, біз тікелей байқалмайтын нысандарға келіп тірелеміз. Бұл жағдайда «тілдік белгі» мен «белгіні білдіруші» өзара психикалық пішін үйлесімінде ғана байланысты болмай, олар мазмұнды тұрғыдағы үнемі өзгерісте болатын экстралингвистикалық ұғымдармен ықтималдық байланыстарға да ұшырайды. Тілдің мазмұндық болмысының толық формалдануына кедергі жасайтын, негізінен алғанда, тілдегі семантикалық құбылыстар.

Осы айтылғандардан мынадай қорытынды жасауға болады.

Адам баласына тән сөйлеу не жазу тілі еш уақытта логикалық робот – компьютерге толық берілуі мүмкін емес. Компьютерлерді тілдің берілу көрінісін анализдеу мен синтездеу және мүмкіндігінше оларды формалдау жағдаяттарында пайдалануға болады. Ал мазмұнға тиісті зерттеулер белгілі бір семантикалық шектен аспайтын көлемдегі формалдануға тиісті жәйтгерде ғана іске асуы мүмкін [1].

Компьютер көмегімен шешілетін жоғарыда сөз болған тілдік көріністерді формалдауға көнетін есептерді квазилингвистика саласына жатқызады. Мұндай есептер сан жағынан аз кездескенімен, олардың қодданбалық шеңбері айтарлықтай мол болады. Квазилингвистикалық есептер адам өмірінің көптеген жақтарын қамтиды. Мысалы, олардың ойдағыдай шешілуі экономикаға байланысты жұмыстарды басқаруда, әскери қорғаныс істерінде және басқа да көптеген ғылыми зерттеу жұмыстарында маңызы зор деп есептелінеді. Бүгінгі тандағы осындай есептердің ең негізгісі ретінде – ақпараттық қор жинау мәселесін айтуға болар еді. Бірақ, бұл мәселені шешудің өзіндік қиындық жақтары бар. Ол, әсіресе, адам баласы

мен компьютердің ара қатынастарын (функцияларын) дұрыс ажыратуға көп байланысты. Мысалы, жиі қайталанатын және оңай формалданатын тілдік бірліктерге қатысты операцияларды: сұрыптау, топтастыру, санау, реттеу және т.б. амалдарға жататын есептерді компьютерге жүктеген жөн. Ал, стандарттауға жатпайтындар – семантикаға қатысты тілдік құбылыстар. Мысалы, мәтіннің мазмұнын ашуға қатысты сөздердің көпмағыналығын айыру, омоним сөздердің мағынасын ажырату және сол сияқты кейбір формалдау процесіне көне бермейтін тілдік құбылыстар бар [1].

Тілдік құбылыстардың біркелкі еместігі, бір қарағанда, компьютерді тіл білімі саласындағы зертгеу жұмыстарында пайдалануға кедергі келтіретін сияқты. Егер біз компьютердің мүмкіндігін толық түрде ескере білсек, тіпті шешімі жоқ-ау деген көптеген мәселелердің де шешімі табылар ма еді. Көбінде, компьютер жадында орындалатын амалдар тілшінің орындайтын формалды іс-әрекетіне ұқсас және олар – көрініс деңгейінде тіл мен сөйлеу аспектілерінің көптеген жақтарын қамти алатыны сөзсіз. Сондықтан, бұл жердегі ең бастысы, адам баласы өзінің ойлау процесін компьютерге бөлшектеп беріп, формалдау жолдарымен ұштастыра білуі қажет сияқты. Сонда ғана зерттеуші тілдік парадигмалық көріністерді, фразалық тармақтанудың заңдылықтарын ашуды және т.б. формалдануға кажетті жұмыстарды іске асырар ма еді [1].

Қазақ тіл біліміндегі зерттеу кезеңдеріне, әсіресе, лингвистикалық ақпараттармен жұмыс істеу, яғни оларды жинақтау, іздеу, сақтау, шет тілдегі ақпараттарды қазақ тіліне аудару, ақпараттық библиографиялық іздеу, қазақ тіліндегі мәтіннен реферат алу және т.б. істерге логикалық роботтың (компьютердің) араласуы жақын болашақта мемлекеттік тұрғыда қайта қаралып жатса, бұл мәселе өндірістік дәрежеге дейін көтерілуі де мүмкін.

Ғылым мен техниканың дамуы тіл қызметінің қоғамдағы коммуникативтік қолданыс аясын кеңейте түскені бәрімізге мәлім. Тіл өзінің адамдар арасындағы карым-қатынастық қызмет атқаруымен бірге адам мен компьютердің қазақ тіліндегі диалогын да іске асыруы қажет деп білеміз. Осы себептен жақын болашақта табиғи тіл және оның түрлері ақпарат таратудағы «адам-компьютер-адам» аралығындағы коммуникативтік жүйе негізін құрайды деп сенім білдіруге әбден болады. Әрине, аталған жүйеде – тіл, өзінің адам баласының ойлау қабілетіндегі үйреншікті ассоциативті-эвристикалық мүмкіншілігінен басқаша, өзіне жат қалыпқа түсіп, компьютерлік тізбектік-алгоритмдік субстратына ауысады. Тілдің бұл құбылысы қазақ тіл біліміндегі ізденістерге біршама методологнялық қиыншылықтар туғызуы мүмкін. Осы себептен, қазіргі тілдік зерттеулердің, біріншіден, қатаң логикалық теорияға бағындырылуын, екіншіден, объективті және толық болуын барынша талап ету қажет [2].

Тіл қызметінің жоғарыда сөз болған шарттары тіл білімінің «инженерлік лингвистика» деп аталатын жаңа саласында қарастырылады.

Инженерлік тіл білімі, негізінен, электрондық машина (компьютер) қажеттігі үшін тілдік белгілерді, тіл жүйесі мен оның атқаратын қызметін (функциясын) модельдеу мәселелерімен айналысады.

Тілдік объектілерді жасанды түрде қайта өндіру үшін тек қана жоғары дәрежедегі техникалық база болуы жеткіліксіз. Біздің ұйғаруымызша, қолданбалы есептер үшін, ең бастысы, күшті және объективті теория құру қажет. Сондықтан да әрбір зерттеушінің тәжірибе жүзінде алған нәтижелері сол лингвистикалық теорияға баға беретіндей өлшемге (критерийге) айналуы керек.

Инженерлік тіл білімінің қажеттігін дәлелдейтін проблемалық бағыт-бағдарлардың ең негізгісі – мазмұн бірлігін автоматты түрде айқындап білу және табиғи тіл мен жасанды тілдерді семиотикалық жағынан барынша жақындастыру проблемалары. Сонымен қатар ашық жүйеге жататын табиғи тіл және келісіммен алынған, жабық семиотикалық жүйедегі қарым-қатынастарды зерттеу – инженерлік тіл білімінің методологиялық және технологиялық мәселелерінің ең бір негізгілері. Бұл мәселенің дұрыс шешім табуы – семиотика, информатика, компьютерлік бағдарлама жасау және математикалық машиналар теориясы, ойлау процестерін модельдеу салаларындағы ғылымдардың қарқындап дамуының бірден-бір шарты [2].

Инженерлік лингвистиканың тіл біліміндегі басқа да пәндер қатарынан тең дәрежеде орын алуы – көптеген тілшілер мен математиктердің, инженер-программистердің жүргізген тәжірибелерінің, теориялық ізденістерінің нәтижесі деп түсінген жөн. Меселен, қазіргі қазақ тілінің «статистикалық лингвистика» және «инженерлік лингвистика» салалары – физика-математика ғылымдарының кандидаты, филология ғылымдарының докторы, профессор Қ.Б.Бектаев есімімен және ол ғалымның шәкірттерінің көптеген ғылыми ізденістерімен тығыз байланысты. Сол сияқты қазақ тілі грамматикасын формалды сипатгауға байланысты соңғы кездегі ғылыми ізденістерімен көзге түсіп жүрген, математик-профессор А.А.Шәріпбаевты және т.б. атауға болар еді. Аталған зерттеу нәтижелері қазақ тілін инженерлік лингвистика әдіс-тәсілдерімен қарастыруда, мемлекеттік тілімізді оқытудың (үйретудің) автоматты түрдегі жүйесін құруда және қазақ тілінің формалды грамматикасының ғылыми негізін құру жолының бастамасы болатынына еш күмәніміз жоқ.

Классикалық лингвистика, көбінде, ақпараттық процестің қандай материаддық ортада жүріп жатқанына мән бере бермейді. Ал, инженерлік тіл білімінде басқаша, яғни табиғи тілдің «адам-компьютер-адам» жүйесіндегі қызметінің ерекшеліктері стратегиялық және технологиялық ортаға тікелей қатысты. Сондықтан тілдік мәліметтерді компьютер арқылы қайта қарастыру жағдайында, адам баласының миы мен компьютер жадының құрылымды-функционалдық сипаттарын салыстыра отырып, оларға тән қасиеттерді ескерген жөн. Олай дейтініміз, тіл білімі саласындағы электрондық машиналар мен адам баласының мүмкіншіліктері жайлы кептеген тілшілер мен математиктер ғылыми әдебиеттер бетінде өз ойларын білдіріп жүрсе де, тіл білімі саласындағы компьютердің болашағы жайлы ойлары әлде де бірізді емес. Мысалы, біреулері ЭВМ-ның жады мен жылдамдығын місе тұтпаса, ал басқалары – компьютер арқылы семантика-синтаксистік, тіпті стильдік ерекшеліктерге қатысты күрделі-күрделі мәселелердің болатынын сөз етіп,

қалай болғанда да компьютерді тілден алшақтатуға тырысатындар да жоқ емес.

Қазір, компьютерлік техниканың барынша дамыған кезінде, оның жадының көлемі, жылдамдығы туралы сөз болмаса да, басқа ғылыми мәселелердің «кезек күттірмейтін» тұстары (мысалы, түрлі сөздіктер түзу) себеп болып, қазақ тілін зерттеуде компьютерлік тіл білімінің жолы болыңқырамай, қарқыны бұрынғыдан да бәсеңсігенге ұқсайды.

Адам мүмкіншілігіне қарағанда, компьютер, сырт кедергіге қарамай, жадындағы ақпаратты әрі тұрақты, әрі берік, әрі ұзақ мерзімге сақтайтыны және қажетті ақпаратты тез арада қағаз не монитор бетіне шығарып бере алатындығы көпшілікке мәлім. Ал, мәтін ішіндегі ақпаратпен (мысалы, мағына мен мазмұн) жұмыс жасауда адамның ми қызметі оның физиологиялық және психикалық қалпына байланысты болатын болса, мұндай қасиет компьютерге жат екені де белгілі [2].

Инженерлік тіл білімінің даму болашағын сөз еткенде, мысалы, жақын он жылдық ішіндегі зерттеудің стратегиясын анықтауда және үйлесімді технологияны таңдауда, тек қана бүгінгі күнгі кибернетикалық техниканың және белгілі шеңбердегі тіл білімі мен нейрофизиологияның ғылыми жетістіктеріне сүйенген жеткіліксіз деп есептейміз. Бұл жерде кибернетикалық техниканың даму қарқынын ескере отырып, кибернетиканың, тіл білімінің және нейрофизиологияның методикалық жақтарын ескеретін кең шеңбердегі, алдын болжай алатын ғылыми көзқарас қажет.

Осы айтылғандарға байланысты мәтіннің туындауы (порождение текста) жағдаятын және оны әртүрлі деңгейде қабылдаудың мүмкіншілігі жайлы қысқаша сөз етелік.

Мәтін – өзара қарым-қатынаста болатын тілдік жүйелердің «норма», «узус» деп аталатын ережелері мен шектеулеріне және тіддің өзіне байланыссыз пайда болатын сырт жағдайларға байланысты туындайды. Мұндағы «сырт жағдай» деп отырғанымыз – мәтінде баяндалатын оқиға мен шектеулерге (узусқа) қатысты контекстік жағдаяттар [3].

Қабылдаушы адамның өзіне бағытталған хабардың мән-мағынасын анықтауы оның тілдік жүйені меңгеру дәрежесіне және тілдің ережелері мен шектеулерін («норма» мен «узус») айыра білуіне, әрі сол хабарламадағы сөз болатын жағдайлармен қаншалықты таныстығына да көп байланысты.

Қабылдаушы адамның мәтіннің мән-мағынасын жан-жақты анықтауы, оның тіддік жүйені және оның «норма», «узус» деп аталатын қасиеттерін толық игеруімен қатар, осы айтылуда пайда болған жағдайдың прагматикалық құндылығын ескере білуіне де қатысы бар.

Енді хабарламаны лингвистикалық тұрғыдан қабылдау деңгейлерін сөз етейік. Мәтінді лингвистикалық тұрғыдан қабылдаудың төменгі сатылары мыналар:

1) грамматикалық деңгей (бұл деңгейде тек сөйлемнің грамматикалық схемасы ғана қабылданады);

2) сөздік деңгей (мұнда сөзтүлғалардың лексикалық мағынасы ғана қабылданады).

3) сөздік деңгейдегі жоғарғы саты – ол қабылдаудың фразеологиялық деңгейі. Бұл қабылдауда сөз тіркестерінің мағынасы ашылады, ал тіркес арасындағы грамматикалық байланыстар ескерілмейді.

Келесі саты – жоғарыда аталған үш түрлі қабылдауды жалпылайтын саты – мәтін бойындағы лексика-грамматикалық деңгейді айқындау. Мәтін бұл сатыда да толық түрде лингвистикалық қабыддаудан өтпейді. Себебі қабылдаушы адам аталған деңгейлерде тек қана сөздер мен сөз тіркестерінің мағынасын ашып, мәтіннің грамматикалық құрылымын анықтағанымен, оның тереңде жатқан семантика-синтаксистік (әрі стильдік) байланыстарын аңғармай қалуы мүмкін. Мұндай байланыстарды анықтау үшін қабылдаушы адам мәтіндегі семантика-синтаксистік деңгейге көтерілуі қажет. Бұл деңгей, әрине, қабылдаушының тілді толық меңгеруін талап етеді.

Қазіргі кезде еуропа тілдеріндегі мәтіндік хабарламалардың грамматикалық, сөздік, фразеологиялық, кейбір лексика-грамматикалық және, тіпті, семантикалық деңгейлерін айқындайтын компьютерге арналып жасалған алгоритмдер негізіндегі бағдарламалары іске қосылған.

Осы айтылғандардың негізінде төмендегідей сұрақ туындауы мүмкін.

Электронды автоматгың (компьютердің) болашақтағы мәтінді түсіндіру мүмкіндігі қай деңгеймен шектелмек? Ол тек лексика-грамматикалық, лингвистикалық деңгей ме? Әлде компьютердің мәтін мазмұнын анықтаудағы мүмкіндігі маман аудармашының деңгейіне теңелуі мүмкін бе? Компьютерлік аударма проблемасына қатысты бұл сұрақты былай да қоюға болады: шетел тіліндегі белгілі бір сала бойынша алынған мәтінді әрі осы саланың маманы, әрі тәжірибелі аудармашы болып табылатын адамның аударуы мен компьютерлік аударманы салыстыруға бола ма? Әлде компьютерлік аударма еш уақытта жоғары сапалы болмай, оған тұтынушының амалсыз араласуы қажет пе? Әрине, қазіргі кездегі компьютерлік аудармалар әлде де маман аудармашының жәрдемін қажет етеді. Соңдықтан да, кибернетикалық автоматтың мәтіңдерді «адамша» аударуы бір адамның бойына біткен лингвистикалық білім дәрежесімен және басқа да табиғи мүмкіншіліктерімен шектелмеуі керек, ол компьютер үшін жеткіліксіз. Болашақтың лингвистикалық компьютері бүкіл бір ұйымға тән, әлеуметтік кәсіби (профессионал) топтарға тән интеллектуалдық, мәдени-әлеуметтік және лингвистикалық тәжірибелерді өзіне игеруі қажет.

Қысқаша айтқаңда, болашақ автомат-роботтың тілдік білім дәрежесі әлеуметтік деңгейде болуы тиіс. Бірақ мұндай дәрежедегі электронды автоматты құрастыру ісі, әзірше, көптеген күтпеген арақайшылықтарға ұшырауда. Осындай қиыншылықтардың негізгісі – адам баласының сөйлеу тіліндегі дискреттік (бөліктенулік) және үздіксіздік тәрізді қасиеттері болып табылады [3].

Мәтінге қатысты инженерлік-лингвистикалық жұмыстар, яғни мәтінді толық қабылдау және оны «тудыру» (порождение) тек «адам-компьютер-адам» (А1-К-А2) коммуникативтік жүйесінде ғана іске асады. Бұл жердегі

адам, яғни «А1» – хабаршы, «А2» – хабарды қабылдаушы рөлінде, ал компьютер яғни «К» – хабаршының хабарламасын қабылдаушы адамға жедел және ыңғайлы түрде жеткізетін аралық компонент.

«Адам-Компьютер-Адам» – хабардың берілу жүйесінің схемасы төмендегідей:

$$A1 \longrightarrow K \longrightarrow A2$$

Суреттегі А1–К–А2 хабардың берілу жүйесін екіге: (А1–К) және (К–А2) бөліп қарастыруға да болады. Бірінші процесте (А1–К) мәтіндік хабарлама адамнан компьютерге беріледі, ал екіншісі – (К–А2), керісінше, – компьютерден адам назарына қарай аударылады. Осы екі түрлі процестің әрбіреуін төмендегідей төрт блокқа бөліп қарастыруға да болады [3].

Бірінші процесс А1–К:

1.1) хабаршының (А1) назарында (ойлау процесінде) хабарлама мәтінінің «туындауы»;

1.2) «туындалған» мәтін белгілерінің (кодтары) импульстер тізбегі ретінде компьютерге берілуі;

1.3) қабылданған импульстер жиынтығының компьютерлік семиозис негізінде мәтінді айқындау және оның мағынасын (мазмұнын) формалдау;

1.4) компьютер жадындағы айқындалған мәтінді мақсатқа сай өңдеу (өзгерту), мысалы, басқа тілге аудару, мәтіннен аннотация, реферат алу, индекстеу және нәтижесін табиғи тіл түрінде қабылдаушыға ұсыну.

Екінші процесс К–А2 төмендегідей төрт блоктан тұрады:

2.1) компьютер жадындағы мақсатқа сай өңделген ақпаратты тұтынушыға ұсыну (яғни бұл блок жоғарыдағы көрсетілген 1.4-ке ұқсас);

2.2) компьютерлік нәтижені қабылдаушы адамның (А2) оқуы;

2.3) компьютерлік хабарламаның семиозисі;

2.4) қабылдаушы адамның (А2) хабарламадан алған әсері.

Хабардың берілу жүйесі қызметінің ең маңыздылары деп, компьютерлік семиозисті (1.3), мәтінді айқындауды және қабылдаушының компьютер ұсынған нәтижеге байланысты түсінігін айтады. Осы аталған жағдайларда да семиозистің жетістігі табиғи тілдің, компьютерлік тілге, яғни жасанды математикалық тілгге ауысу нәтижесінің дәрежесіне тығыз байланысты.

Инженерлік тіл білімінің келесі бір маңызды проблемасы – лингвистикалық алгоритмдер мен олардың программаларын құру мәселесі. Мұны басқа сөзбен, «компьютерді оқыту» деп те айтуға болады. Табиғи тіл мен жасанды тілдің қатынастық жақтарын дұрыс түсінбейінше, бұл мәселенің дұрыс шешіміне қол жеткізу қиын.

Табиғи тіл де, жасанды математикалық (компьютерлік) тіл де семиотикалық жүйеге жататындықтан, олардың негізгі қасиеттеріне тоқталып өтейік.

Тіл жүйесі деп тілдік бірліктерді және олардың арасындағы қатынастарын айтамыз. Ал табиғи тіддің бірліктеріне жататындар: дыбыс пен фонема, әріп, морфа мен морфема, сөз бен сөзтұлга және олардың құрылымдық түрлері, сөз тіркесі, сөйлем және олардың қаңқалық схемасы.

Енді жасанды компьютерлік тілдің бірліктері – әріп, цифрлар мен символдар және олардың шартты белгілері (кодтары), компьютерлік түбір сөз, сөздер және сөйлемдер. Бұл жердегі компьютерлік сөз не түбір сөз деп отырғанымыз, белгілі бір түрде шектелген әріп тіркестері, ал компьютерлік сөз тіркес және сөйлемдер дегендеріміз, алдын ала келісім бойынша шектелген компьютерлік сөздер тізбегі деп ұққан дұрыс.

Табиғи тілдің мынадай үш түрлі функцияларын атауға болады [4]:

а) семантикалық және сигматикалық ақпаратты таратуға мүмкіндік туғызатын лингвистикалық символды қалыптастыратын тілдің номинативті функциясы;

б) тіддік белгілерді өзара байланыста ұстайтын және мәтіндік ақпараттың синтактикалық қасиетін сипаттайтын тілдің предикативті функциясы;

в) хабардың берілу кезінде тілдің номинативті және предикативті функцияларын хабарлаушы (А1) не қабылдаушы (А2) жеке адамның (жеке ұйымның) қалауымен ұштастыра алатын прагматикалық функция.

Жасанды математика тілінде және оның барлық тармақтарында, яғни компьютерлік тілдің берілуінің өзінде-ақ предикативті функция біте қайнасып жатады.

Жасаңды компьютерлік тілдің номинативті функциясы белгінің (знактың) мағынасымен (десигнатымен) шектеледі. Ал, бұл жасанды математикалық тілде прагматикалық функция атымен жоқ деуге болады [3].

Сонымен, хабардың берілу жүйесі (А1-К-А2) жұмысының нәтижелігі, біріншіден, мазмұнды хабардың көлемдік мөлшеріне және екіншіден, көпшілік қабылдаушының дайындық дәрежесіне де көп байланысты демекпіз.

Қорыта келе. жоғарыда сөз болған «адам-компьютер-адам», яғни хабардың берілу жүйесінің салалық тілдермен жұмыс істеуі, табиғи тіліміздің коммуникативтік қызметін дамыта түсетіндігіне еш күмән жоқ деуге болады. Ал, қазақ тіл білімінің инженерлік саласының тиісті дәрежеде дамуың, оның қатаң ережеге бағынуынан, мәтін бойындағы тілдік бірліктерді тануға қажетті формалды қасиеттерінің айқындалуынан, омоним сөздер мен көпмағыналы сөздердің және синоним сөздер қатарын сирету мен реттеу мәселелерінің өз шешімін дұрыс табуынан байқауға болар еді.

Сонымен, инженерлік тіл білімінің жалпы тіл теориясына қатысты атқаратын кейбір қызметтері осындай.

Әдебиеттер

1. *Андреев Н.Д.* Язык в обществе, использующие кибернетические машины // Вопросы социальной лингвистики. Л.: Наука, 1969. С. 25-40.

2. *Пиотровский Р.Г.* Экстралингвистические и внутриязыковые вопросы при переработке текста в системе «Человек-машина-человек» // Вопросы социальной лингвистики. Л.: Наука, 1969. С. 40-63.

3. *Пиотровский Р.Г.* Текст, машина, человек. Л.: Наука, 1975. 327 с.

4. *Баранов А.Н.* Введение в прикладную лингвистику. М.: Эдиториал УРСС, 2001. 360 с.

АБЗАЦ – МӘТІН ҚҰРЫЛЫМЫНЫҢ НЕГІЗГІ СЕМАНТИКАЛЫҚ КОМПОНЕНТІ

Әл-Фараби атындағы Қазақ Ұлттық университеті.
«Хабаршы» журналының филология сериясы,
№ 4 (76). Алматы-2004, 60-63-бб.

Қазіргі кезде егеменді елімізде қазақ тілінің мемлекеттік тіл ретінде дамуына зор мән берілуі тілші ғалымдардың өз ісіне жаңаша көзқараспен қарауын талап етеді. Оған ең негізгі себеп – тілді компьютерлік технологияның жаңа жетістіктерін пайдалана отырып зерттеудің, оның нәтижелерін дұрыс тұтынудың қажеттігі. Міне, осыған байланысты қазақ мәтіндерінің өзіне тән жасалу немесе туындау моделін құру қазіргі заманға сай, кезек күттірмейтін мақсаттардың бірі болмақ.

Қазақ тілінде көптеген іс қағаздарының, неше түрлі хабарламалардың, ғылыми-техникалық әдебиеттер мен мектеп және жоғары оқу орындарына арналған оқулықтардың жарық көре бастағаны бәрімізге мәлім. Олай болса, қазақ тіліндегі мәтіндерді компьютер көмегімен автоматты түрде өңдеу мен қажеттілікке сай мақсаты тіл зерттеу тәжірибесіндегі жаңа сала – мәтін теориясымен тығыз шұғылдануды керек етеді. Жедел дами бастаған мәтін теориясы көптеген күрделі проблемаларды кеңінен қарастыруға жол ашты. Ұлттық тіліміз үшін бұл тұрғыдан мәтіннің негізгі бірліктерінің мәнін ашу және олардың өзара байланысының сыр-сипатын, тақырып рөлін ашу мәселелері аса маңызды. Дәлірек айтқанда, оқырманның көлемді мәтіндердің мағынасын аз мезгілде дұрыс ұғына білуі, оқулықтар мен оқу құралдарын жазу кезінде мәтін құрылымының заңдылығын білуі қажет. Міне, сондықтан да қазақ мәтіндерінің ішкі заңдылықтарын егжей-тегжейіне дейін зерттеу – заман талабы десек артық айтқандық болмайды. Бірақ, бұндай зерттеудің өзіндік қиыншылықтары жоқ емес. Осыдан біраз жылдар бұрын мәтін ұғымының ешбір анықтамасыз қыбалданып, ал тіпті берілген анықтамалардың өздері де бірізді болмай жүргені, бұл істің оңайға түспейтіндігіне дәлер болар деп ойлаймыз. Олардың кейбіреулеріне мысал келтірерлік: «мәтін дегеніміз ұзын сөйлем»; «мәтін – семантикалық жағынан ұйымдасқан белгілер жүйесі» және т.б. Ал, логика-лингвистикалық тұрғыдан құрылған анықтама бойынша мәтіннің басты белгісі ретінде сызықтық тізбе құрайтын сөйлемдерден тұратындығы және олардың сол жақтық, оң жақтық шекарасы болатындығы, салыстырмалы толықтық пен байланыстық шарттарының орындалуы тәрізді ерекшеліктердің басшылыққа алынуы байқалады.

Осындай анықтамалар негізінде біз мәтіннің мынадай негізгі қасиеттерін атап кеткенді жөн көріп отырмыз. Ол – мәтіннің бір мазмұнды

болуы және оның бірліктерінің өзара грамматикалық, семантикалық әрі логикалық байланысқа түсуінің қажеттігі.

Мәтіндерді өзара байланысты немесе байланыссыз деп ажыратудың негіздеріне сүйене отыра, соңғы жылдарғы зерттеулерде оның сыр-сипатын ашуға арналған талдау (анализдеу) әдісінен гөрі, жинақтау (синтездеу) әдісінің көбірек қолданылып жүргенін байқауға болады. Байланысты мәтіндер жайлы теория, әсіресе сөйлемдердің өзара тіркесу заңдарының синтаксистік тұрғыдан баяндалған нақты ережелері көрсетілген зерттеулер, әлі де болса жетіспейді.

Отандас тілші ғалымдарымыздың дәстүрлі лингвистикалық зерттеулері де жеке сөйлем шеңберінен тыс шыға қоймағандығы бәрімізге мәлім. Сондықтан, мәтіннің тек қана сөйлемдер мағынасымен шектелмейтінін, одан да ірі мәтін бөлшектерінің белгілі реттілікті сақтайтын фразадан жоғары (сверхфраза) бөліктерімен біртұтас мағына түзейтінін ескерту қажет болып отыр [Гальперин 1981]. Психологтардан кейін көп ұзамай-ақ тілшілердің өздері де мәтін бойындағы жалпы тақырыптың подтема, субподтема, микротема және т.б. терминдермен аталатын шағын тақырыпшаларға ажыратылатынын қолдай бастағаны белгілі [Новиков 1983; Маслов 1975].

Тіл жүйесінде осы күнге дейін толық шешімі табылмай жүрген мәселелердің ұзын-ырғасы аз емес. Солардың бірі – сөйлемнің жалпы тіл жүйесіндегі алатын орны, екіншісі – сөйлем жасаудың семантикалық және синтаксистік принциптері. Сол сияқты, сөйлемді мәтін ішіндегі негізгі семантикалық компонентке жатқызу не жатқызбау мәселесі де өз шешімін таппай келе жатқан жайдың бірі [Зубов 1987].

Мәтін бойындағы абзац мәселесі де, қазақ тілшілерін айтпағанның өзінде, шетелдік тіл мамандары арасында да әлі де көп талас тудырып жүргені мәлім. Осы кезге дейін абзацты мәтіннің қай бірлігіне жатқызамыз деген сұраққа әр түрлі пікірлер айтылып жүр. Ғалымдардың кейбіреулері оны композициялық (құрылымдық) бірлік десе, екіншісі біреулері – логикалық не семантикалық топқа қосып жүр. Тағы бір зерттеушілердің айтуынша абзац – бір орынға жай топталған сөйлемдердің жиынтығы емес, ол автордың сөйлеу практикасында өз ойын оқырманның саналы түрде қыбалдануына, оны меңгерудің амалдарын жеңілдетудің мақсатына лайықтап құрған мәтіндік бірлік. Мысалы, Г.В.Бондарко, Ю.А.Шрейдерлердің айтуынша, мәтінді абзацтарға бөліктеу – қабылдаушының мәтін мазмұнын дұрыс түсініп, есте қалдыру мүмкіндігін жеңілдету үшін жасалатын арнайы әрекеті екен [Бондарко, Шрейдер 1978]. Біз де осы соңғы пікірді ұстанатынымызды айта кетпекпіз.

Қоршаған табиғи ортаның әр түрлі нақты нысандардан тұратындығы және оларға тән динамикалық, статикалық, құрылымдық қасиеттері болатындығы, кеңістікке тән мезгілдік қарым-қатынас қасиеттерін иеленетін жиындардан тұратыны айқын. Мұндай ортаны танып білу дегеніміз, сол ортаны мүмкіндігінше кіші бөліктерге бөліп қарастыру, яғни ол біздің санамызда немесе мәтін ішінде кездесетін өзара мағыналық байланыстағы сөздер бірлігінің (блгының) көрінісі екен [Зубов 1985: 34]. Сондықтан да біз,

көптеген тілшілердің, психолотар мен логика мамандарының, сонымен қатар өз зерттеулеріміздің негізінде, аталған өзара мағыналық байланыста тұратын сөздер блогы – жазба мәтін, көп жағдайда, абзацпен бірдей болады деп ұйғаруды жөн көрдік. Жазба мәтіннің абзацтардан құралған бөліктерге бөлшектелу себебін анықтау үшін ғалымдар әр түрлі тәжірибелер жүргізіп, олардың кездейсоқ жағдай емес екендігін, оның мәтін иесінің мазмұнды саналы түрде «тізгіндеу» қажеттігінен туындайтынын дәлелдеді.

Абзацқа енетін мәтіннен алынған сөйлемдерді өзара байланысы жоқ бірнеше абзацтардан алынған сөйлемдер тізбегімен салыстыра қарастырғанда төмендегідей ерекшеліктерді байқауға болады:

а) абзац құрайтын сөйлемдер тізбегі *бір микротеманың* не болмаса бір-бірімен тығыз байланыста тұрған *бірнеше микротеманың* мазмұнын анықтайды;

ә) бір абзацтың ішіндегі сөйлемдер өзара арнаулы лексикалық амал-тәсілдер арқылы байланыста болуы керек,

б) абзац ішіндегі сөйлемдердің бір-бірімен белгілі синтаксистік байланыста болуы да шарт [Агроскина 1972] .

Ғылыми-техникалық мәтіндерде бұл ережелер тұрақты қасиетке ие, яғни стереотипті сипатта, ал көркем әдебиет стиліндегі абзацтар көрінісі біршама еркін, яғни шығармашылыққа қатысты өзгеріп отыруы, әр авторға тән өзіндік айырым-белгілері болуы мүмкін. Сол сияқты, ғалымдар мәтіннің семантикалық құбылуын абзацтың құрылымдық (композициялық) ерекшелігімен де байланыстырады.

Біздің қарастырғанымыз мәтін құрылымындағы абзацтардың атқаратын қызметіне тән белгі – стереотиптік (қалыптық) қасиет. Күнделікті өмірімізде, яғни білім алу кезеңінде не тіл үйрену жағдайында біздің сана-сезімімізде әр стильге, яғни ғылыми, публицистикалық көркем т.б. әдебиеттердің негізгі белгі-қасиеттері тұрақты стереотиптік түрде жинақты орын алады. Аталған қасиеттер жиынтығын мәтіннің синтаксистік құрылымының ықтималдық моделі деуге болады.

Сөйлеу қызметінің «астарында» сөйленіс мазмұнында оның түп қазығы ретіндегі заттық мағына жатқаны айқын. Мұндай тұжырым бойынша тілдік коммуникацияның (қарым-қатынастың) негізі дискретті, яғни бөліктік тұрпатта болатындығын, олай болса оны сәулелендіретін мәтіннің де бөліктелген «кванты» немесе «бөлімшесі» – «логикалық қалып» (логикалық стереотип) болады деп ұйғарған жөн [Герд 1978; Серкова 1981]. Сол сияқты көптеген ғылыми ізденістерде ми қызметінің формасын (тұрпатын) элементарлық ақпараттар процесінің жиынтығы деп қарастыру жиі кездеседі.

Сонымен, мәтін бойындағы абзац «мағыналық квант» пен «қатып қалған» синтаксистік үлгі ретіндегі мәтіннің ең кіші логика-семантикалық компоненті (бөлігі) деп қарастыруға әбден болады.

Орыс тілі және басқа да үнді-еуропа тілдеріндегі зерттеулерді негіз етіп алып, біз қазақ тіліндегі мәтіндердің стильдік айырымдары бойынша ажыратылған абзацарға тән функционалды-мағыналық топтарының статистикалық ерекшеліктеріне қысқаша тоқталамыз [Жубанов 2002].

Барлық тілдердің, соның ішінде қазақ тілінің мәтіндерінің айрықша қасиеті – болмыстағы нысандар мен құбылыстардың қарым-қатынастарын сипаттайтын оқиғаны баяндайтын сөйленіс түрінің – сөйлемдердің бір-бірімен тығыз мағыналық байланыста болуы. Ал, бұл болмыстағы кейбір заттар мен құбылыстар тізбектеріндегі, уақыт не кеңістік арасында пайда болатын қатынастың мағыналық тұтастыққа ие болатындығының айғағы [Мильчин 1980: 112].

Ғалымдардың тұжырымдауынша, мәтін бойындағы сөйлемдерде айтылымдар мен сөйлеу типтері *суреттеу, баяндау, пайымдау* түрінде қосылыс табады. Ал, олардың әрбіреуі белгілі бір функционалды-мағыналық мәтіндік және тұрақты дәрежедегі құрылымдық сипатқа ие.

Енді осы аталғандарды сөйлеу типтеріне, яғни *баяндау, суреттеу* және *пайымдау* түрлеріне ажырата білу үшін олардың анықтамаларына тоқталайық.

1. Егер мәтін ішіндегі фразалардың немесе одан ірі мәтіндік бірліктердің орналасу тәртібі оқиға желісінің уақыттық деңгейіне қатысты болса, онда сөйлеу түрін *баяндау* типіне жатқызу қажет.

Баяндау типі іштей етістіктің семантикасына қарай, *нақты-оқиғалық* және *жалпылама-оқиғилық* болып ажыратылады. Ал, мәтіндегі хабарлар мен жалпы ақпараттық баяндаулар тізбегі үшін кейбір қимыл-әрекеттерді немесе тұрақтылық күйді баяндауда мезгілдік нақтылықтың қажет болмауы да мүмкін.

2. Егер мәтін бойындағы фразалар немесе одан ірі мәтіндік бірліктер кеңістік, функционалдық не басқа деңгейдегі қатынаста болатын болса, ондай сөйлеу түрі *суреттеу* типі деп аталады.

Оқиға желісіндегі *суреттеу* типін ажыратудың ең маңыздылары төменде көрсетілетін құрылымдық-мағыналық түрлер [Жубанов 2002: 117]:

1) табиғатты суреттеудің түрлері;

2) ғимараттың ішкі құрылысын суреттеу;

3) кейбір жанды не жансыз нысанның сыртқы пішінін суреттеу.

Мысалы, адамның сыртқы түрін, киімін, жан-жануардың, өсімдіктің, ғимараттың т.б. сыртқы пішінін суреттеу;

4) кейбір тірі жәндікті (адам, жануарлар) суреттеу немесе жансыз заттарға берілетін суреттеме-мінездемелер. Мысалы, адам баласының мінез-құлқы, қызмет түрі, оның икемдік қасиеттері және т.б.

3) Егер фразалар немесе одан ірі мәтіндік бірліктер себеп-салдар деңгейіндегі қатынаста болса, ондай сөйлеу түрі *пайымдау* типі деп аталады.

Функционалды-мағыналық сөйлеу типі ретіндегі *пайымдаудың* түрін логикалық тұрғыдағы ой қорытындылау (ой түйіндеу) дәрежесінде түсіну қажет, яғни логикалық ойлаудың ең жоғарғы формасы деп түсінген жөн. Басқаша айтсақ, *пайымдау* типі логикалық категорияға қатысты бола тұра, мәтіннің белгіл-бір құрылымды-мағыналық бірлігі болып та есептеледі [Жубанов 2002: 118].

Қазақ тіліндегі тәжірибе ретінде алынып отырған ғылыми-көпшілікке арналған мәтіндерді зерттеу негізінде пайымдаудың алты түрі және оларға тән тілдік тұрпат (форма) анықталды.

Тәжірибелік мәтін бойынша компьютер арқылы жүргізілген статистикалық есептеу негізінде ең жиі қолданыстағы *пайымдау* түрлері мыналар:

1) есім түріндегі бағалау, пайымдау түрі;

2) тұрақтылық күй мағанасындағы пайымдаулар;

3) нақты немесе жорамал іс-әрекетті негіздеу мағынасындағы пайымдаулар;

4) кейбір іс-әрекеттің қажеттігін, мүмкіндігін білдіретін немесе қалау не тілеу мағынасындағы пайымдаулар;

5) алдын-ала келісілген іс-әрекет мағынасындағы пайымдаулар;

6) кесімді түрдегі терістеу не тұжырымдау мағынасындағы пайымдаулар.

Мұндай *баяндау, суреттеу, пайымдау* типтерінің айырымдық ерекшеліктері мәтін ішінде ең кіші семантика-синтаксистік бірлік (единица) арқылы, яғни абзац түрінде шектеліп барып, көрініс табады. Сондықтан, біз өз зерттеуімізде мәтіндегі абзацтарды – *баяндау абзацы, суреттеу абзацы* және *пайымдау абзацы* деп сараладық. Кейбір абзацтарға сөйлеу типтерінің аталған үш түрінің де қасиеттерінің қатысы болып жатса, ондайды *аралас типті* абзацтар деп атадық.

Кез-келген мәтінде (ғылыми стильден басқа) сөйлеудің диалог түрі кездеседі, сондықтан сөйлеуге қатысушы екі жақтың да айтылымдарын жоғарыда аталған – *баяндау, суреттеу, пайымдау* абзацтары түріне қосалқы тағы да мынадай функционалды-мағыналық түрлер ажыратылды: а) *хабарлау* мәнді абзацтар: ә) *сұрақ* мәнді абзацтар; б) *әсер ету* (лепті мағыналы) абзацтар; в) *ниет, ойды* білдіретін абзацтар; г) *пікір тудыру* абзацтар; д) *аралас* абзацтар [Жубанов 2002: 118-119].

Осы аталғандардан басқа, біз өз зерттеуімізде абзацтар қызметіне, орналасу тәртібіне қарай: мәтіннің бастапқы жағында орналасқан – *кіріспе* абзацтарды, ішкі бөлігінде кездесетін *медиалды* абзацтарды және мәтіндегі оқиға желісін аяқтайтын – *қорытынды* абзацтарды ажыраттық [Жубанов 2002: 119].

Зерттеу нысаны ретінде үш түрлі стиль бойынша: газет мәтіні, ғылыми-көпшілікке арналған мәтіндер, жеке көркем әдебиет мәтіндері қарастырылды және оларда кездескен абзацтарды жоғарыда аталған түрлер бойынша топтап, компьютер көмегімен алынған статистикалық деректерге салыстырмалы талдау жүргізілді. Енді сол бір тәжірибелік сипаттағы нәтижелердің кейбіреулеріне қысқаша тоқталайық.

Көркем әдебиет бойынша 487 абзацтан тұратын 20 түрлі мәтін қарастырылған болатын. Алынған статистикалық нәтижелер арнайы кесте түрінде көрініс тапты. Мақалада сол кестелерден алынған кейбір деректерді сөз етпекпіз.

Қарастырылған мәтін бойынша *кіріспе* абзацтар, *медиалды* (ішкі) және *қорытынды* абзацтар функционалды-мағыналық жағынан сараланып, олардың қандай типтері өнімді, қандай типтері өнімсіз екені сараланып, нәтижесін көркем әдебиетке қатысты мәтіндер бойынша айқындалған 44 *кіріспе* абзацтардың 30 (68%) *баяндау* мәнді абзацтары екендігі анықталды. Сол сияқты, абзацтардың 20,4 пайызы *аралас* типта абзацтарға жататынына көз жеткіздік. Ал, *кіріспе* абзацтарда функционалды-мағыналық типтердің басқа түрлері, яғни *суреттеу* мен *пайымдау* түрлері аз қолданыс тапқаны анықталды.

Көркем әдебиет мәтіні бойынша, *баяндау* типті абзацтардың *ішкі* (*медиальный*) және *қорытынды* абзацтар ішінде де көп қолданыста болуын (54% және 69%) статистика жолымен дәлелдедік.

Сан жағынан 404 *медиалды* (*ішкі)* абзацтар қатарынан жоғарыда аты аталған типтердің барлық түрлері де (біреуі аздау, біреуі көптеу) орын тапты деуге болады. Олардың ішінде бірінші орында – *баяндау*, одан кейінгі екінші орында – *аралас* типті, сосын барып, жиілігінің кему ретімен алғанда: *әсер ету, сұрақ мәнді пайымдау* және *суреттеу мәнді* абзацтар саралау санағынан өтіп, олардың типтері анықталды. Бұларда ең жиі кездесетіні *баяндау* мәнді абзацы болғанымен (69%) әрі қарай бөлінуде басқашалау, яғни екінші орында – «*әсер ету*» (15,4%), үшінші– «*пайымдау*», ең соңында «*аралас мәнді*» абзацтар байқалады.

Газет мәтіні бойынша шамалас көлемдегі, тақырыптас («Алты жастағылар мектепке барады») 35 мақала зерттелді. *Бастапқы* және *қорытындылау* абзацтары негізінен «*баяндау*» абзацтары (89%, 96%) болса, ішкі (*ортаңғы, медиалды*) абзацтар ішінде ен көбі – *баяндау мәнді* абзац, сосын *аралас мәнді* (17%), *суреттеу* (9%), *сұрақ мәнді* (2%) және *әсер ету мәнді* (1%) абзацтар типтері.

Ғылыми-көпшілікке арналған қазақ мәтіндерінің 47-сі қарастырылды. Бұл аталған функционалды тип мәтінінде де жоғарыда аталған жалпы заңдылық, яғни *баяндау мәнді* абзацтарының (*кіріспе, ішкі, қорытынды*) басылымдылығы сақталады. Дәлірек айтсақ, *бастапқы* және *қорытынды* абзацтардың 70 пайызы, *ішкі* абзацтардың 50 пайызы *баяндау типті* болып кездескен. Бұл жерде айта кететін жайт, *бастапқы* абзацтардың 17 пайызы *аралас* типті, бірақ олар негізінен *баяндау* мен *пайымдау* типтеріне ұқсас келеді. Мұндай *аралас* типтердің басымдылығы мәтіннің *ортанғы* және *соңғы* бөліктеріне де тән екені айқындалды. Ал, *ішкі* абзацтар тобына жататындарда *баяндау* типінен кейін екінші орында *пайымдау мәнді* абзац екені анықталды. Ол барлық *медиалды* абзацтардың 28,4 пайызын құрайды. Тағы да бір айта кететін жайт, ғылыми-көпшілікке арналған қазақ тіліндегі мәтіндердің «шеткілерінде», яғни «кіріспе» және «қорытындыларында» *пайымдау* типі үшінші орынға ауысып, 5% және 13% шамасында кездесті.

Бұл жүргізілген тәжірибелік зерттеуіміздің негізінде абзац типтерінің біз көрсеткен пайыздық дәрежелері осы аталған стильдік мәтіндерге әр уақытта тән қасиет деуден аулақпыз. Себебі, бізге қиындық келтіріп отырған жағдай – абзац түрлерін әр түрлі типтерге жатқызудың айырымдық жағын

жете айыра білу, яғни олардың шекарасын ажырата білу. Бұл жерде зерттеушінің сол мәтінде сөз болатын тақырыпқа мамандық жағынан қатынасы болуы қажет сияқты және мәтінге тән басқа да өзіндік айырым-ерекшелікке байланысты статистикалық нәтижелер сандық-сапалық өзгерістерге ұшырауы мүмкін деп ойлаймыз.

Бұл мақалада қазақ мәтіндерінің құрылымындағы абзацтардың функционалды-мағыналық топтарына қысқаша тоқталдық. Ал осыны нақтылау тұрғысынан абзац түрлерінің мазмұндық жағынан заттық-логикалық топталуы да мәтіннің «туындау» мәселесінде аса зор орын алатынын ескертпекпіз. Жалпы, «Қазақ тілінің мәтін туындату алгоритмін» құру мәселесінде мынадай мақсат пен міндеттер ең алдымен шешілгені жөн деп есептейміз:

1) формалды түрде мәтіннің жалпы тақырыбының берілуі;

2) мәтіннің негізгі семантика-синтаксистік бірліктерінің айқындалуы;

3) осы бірліктерді тұтастық байланыста ұстайтын ықтималды-анықталған ережелердің берілуі [Жубанов 2002].

Бұл қазақ тіліндегі әр түрлі құрылымдағы және көлемдегі мәтін түрлері мен типтерін арнайы тәсілдермен терең зерттеуді қажет етеді.

Әдебиет

1. Гальперин И.Р. Текст как объект лингвистического исследования. – М: Наука, 1981. – 139 с.

2. Новиков А.И. Семантика текста и ее формализаиия. – М : Наука. 1983. – 261 с

3. Маслов Б. А. Проблемы лингвистического анализа связного текста (межфразовый уровень), Учеб. пособие к спецкурсу. – Таллин: 1975. –104 .с

4. Зубов А.В. Проблемы порождения текста Часть 1. Теория и алгоритмы. – Минск, 1987. – 77 с.

5. Бондарко Г.В., Шрейдер Ю.А. Текст–смысл–ситуация. – В кн.: Вопросы информационный теории и практики. № 36 – М.: ВИНИТИ, 1978 – С. 80-91.

6. Зубов А.В. Вероятностно-алгоритмическая молель порождения текста (семантика-синтаксический аспект). – Минск: 1985 – 399 с. (докт. дисс. филол. наук).

7. Агроскина С. Н. Структура абзаца в художественной прозе. Автореф. дис. канд. – Л., 1972 – 25 с.

8. Герд А. С. Логико-понятийная структура научного текста и его семантика. – В кн.: Вопросы семантики Калининград. 1978. – С. 55-58.

9. Серкова Н.И. Семантики и метасемантика красной строки в художественной прозе. – В кн.: Актуальные вопросы лингвистического и методологического изучения художественного текста. - Владимир: 1981. – С. 108-120.

10. Жубанов А. К. Основные принципы формализации содержания казахского текста. – Алматы, 2002. – 250 с.

11. Мильчин А.Э. Методика редактирования текста – М.: Книга, 1980. – 318 с.

ТІЛДІ ЛОГИКАЛЫҚ ТАЛДАУ БАҒЫТЫНЫҢ КЕЙБІР МӘСЕЛЕЛЕРІ

Әл-Фараби атындағы Қазақ Ұлттық университеті.
«Хабаршы» журналының филология сериясы,
№ 5 (77). Алматы-2004, 37-43-бб.

1986 жылы Ресей ғылым академиясының Тіл білімі институтында «Логический анализ языка» (ЛАЯз – «Тілді логикалық талдау» – ТЛТ) теориясымен айналысатын ғылыми топ құрылған болатын. Аталған теориялық бағыттың ғылыми жетекшісі – филология ғылымының докторы, Ресей ғылым академиясының корреспондент мүшесі Н.Д.Арутюнова, негізінен, тілді концептуалды тұрғыда қарастырудың әдістері мен категорияларындағы ойлау мен білімге қатысты логикалық талдаулар жасау ісіне аса көңіл бөледі.

Тілді зерттеуде неше түрлі формалды әдістемелік бағыттар қолданылатыны мәлім. Олар: құрылымдық және математикалық талдау әдісі, дискриптивті және генеративті лингвистика тәсілдері, тілді «мағынадан мәтінге қарай» сипаттау моделі, тілді дистрибутивтік және компоненттік талдау, функционалды грамматика, прагматикалық әдіс және т.б. Сол сияқты тілді компьютер арқылы зерттеуде, әсіресе мәтінді автоматты түрде талдауда практика мен теорияның ара жігін айыру мүмкін еместігі анықталды. Тілді зерттеуде логикалық әдіске сүйену ең алдымен адам баласының ойлау жүйесінің сырын ашу қажеттігінен туындайды. Мұндай мәселенің шешімі тек табиғи тіл арқылы ғана анықталатыны мәлім болды. Міне, сондықтан зерттеудегі тілдің логикалық негіздемесіне бетбұрыс жасау осы саладағы кейбір әдістемелік және концептуалдық қиыншылықтарды жеңу мен тілдің ішкі табиғатын терең түсінуді көздейді.

Айтылым кезінде ойлау категориясы, ғаламның тілдік бейнесі мен аялық білімі жайлы, өмір сүру логикасы мен пайымдау тәжірибесі, айтушының ішкі дүниесіне әсер ететін психологиялық механизмдер (саналы не санасыз түрдегі), тілден сырт болмыс, ашық не жасырын түрдегі хабар мақсатының тікелей коммуникативтік жағдаяты (иллокутивтік күші) және т.б. факторлардың қатысатыны белгілі. Осының негізінде «Тілді логикалық талдау» арқылы зерттеудің аясына прагматикалық категорияларды да қосып алуға тура келеді. Бұл жағдай тілдегі семантикалық аппарат арқылы нақты сөздердің мағынасын ашуды ғана емес, әр қилы концептердің де мағынасын қоса зерттеуді кажет етеді. Мұндағы концепт ұғымын «адамның әлем туралы жинақтаған мәдени түсініктерін бейнелейтін атаулардың жүйесінен тұратын ғаламның тілдік бейнесі ретінде» тануға болады [23, 85-б.]. Оны тілімізде

қалыптасатын және ұрпақтан-ұрпаққа өзгеріссіз беріліп отыратын тілдік қор деуге де болады.

Жоғарыда аталған ТЛТ ғылыми бағдарламасында осындай бірқатар проблемалық мәселелерге ерекше көңіл бөлінді. Мәселен, олар менталды перцептивті етістіктердің ***(знать, видеть, слышать, полагать, верить*** и др.) айтылым мағынасының пайымдау ақиқаттығына әсерін тигізетін пропозицияға қатынасын қарастырумен де шұғылданды: ***Я думаю (верю, полагаю, знаю, сомневаюсь), что ты говоришь правду.*** Аталған етістіктердің қазақ тіліндегі баламалары: ***білу – білемін, көру – көремін, есту – естимін, ойлау – ойлаймын, сену – сенемін,*** т.б.***; Сенің айтқаның шындық деп ойлаймын (сенемін, білемін, ұйғарамын).*** Айтушының (субъект установки) пайымдау ақиқаттығына қатысын білдіретін пропозиционалды ұстаным (установка) түрлерін анықтауды қажет етеді. Мысалы, оған ұстанымдарды категорияларға таптастыруды (менталды, сенсорлы, перцептивті, волитивті, прескретивті және т.б.) жатқызуға болады. Осы ұстанымдардың әр типінің пайымдаулармен қарым-қатынасын, басқа адамның сөзін жеткізу кезіндегі айтушының пікірі мен ұстаным субъектісінің пікірі арасындағы қатынасты, терістеудің аясын ұлғайту үшін тәуелді пропозицияға сұрау есімдігінің енгізілуін (***Кім келгенін мен білем; Кім келді деп ойласам***), тәуелді пропозицияның рай, шақ категорияларына қатынасын (***оның кеткені бізге белгілі; оның кеткенін біз білеміз***) және т.б. мәселелерді анықтау мақсатында көптеген зерттеулер жүргізілді. Қысқаша айтқанда, аталған ТЛТ ғылыми тобының алғашқы зерттеулері осындай логика-прагматикалық аспектісіне арналған [1-6].

Тілдің логикалық аспетісіне арналған соңғы онжылдықтағы зерттеулер мәдени концепті мәселесін талдауға қарай бет бұрды. Адам баласының өзі жасаған мәдениеті – **«екінші шындық»** енді кешенді түрде зерттеуді қажет ететін танымдық нысан болып алды. Мәдениет өзін туындатқан халықпен біте қайнасып жататыны айқын және оның бойында – ***ақиқаттық, шындық, жалғандық, бостандық, тағдыр, зиянкестік, жақсылық, сұлулық, тәртіпсіздік, адамгершілік, заң, тәртіп, мақсат, айып, кінә*** және т.б. сияқты өмірлік философияны анықтайтын, барлық адамзатқа тән әлемдік бейнедегі ұғымдардың берік орын алатыны да белгілі. Солай бола тұра, осы аталған ұғымдар ұлттық дүниетаным, мәдениет негізінде қалыптаса отырып, ұлттың ұлт болып айқындалу белгісі ретінде де қолданылады. Бұлардың айырықша мағыналары тек мәнмәтіндік (контекстік) қолданылу аясында ғана ажыратылады, ал ол (мәнмәтін) шартты түрде қайсы бір концептінің **«тілі»** немесе **«грамматикасы»** деп аталады. Сондықтан да қазіргі философиялық мектептердің (феноменология, лингвистикалық философия, герменевтика) және басқа да бағыттар әрдайым тілдік деректерге жүгінуді дәстүрге айналдырып отыр. Шынында да, сөздердің шығу төркіні (этимологиясы), олардың тіркесімдік қабілеті, синтаксистік құрылым типтері, семантикалық өрісі, бағалауыштық қасиеті, ассоциациялық және метафоралық бейнеленуі және т.б. ұғымдар үшін ерекше «тіл» туындайды деуге болады. Бұл жағдай ұлттық болмыс, таным ерекшеліктерінің адам санасындағы орнын

анықтайтын концептіні қайта құруға мүмкіндік жасайды. Мәдениетке қатысты концепті адам мен ол өмір сүретін ақиқат шындық аралығын байланыстырушы ретінде қызмет атқаратындықтан оны зерттеудің маңызы зор.

Логикалық және логика-прагматикалық талдаумен қатар, концептуалды талдау арқылы зерттеудің екінші бағытына «түрлі тілдер мен халық мәдениетіндегі **«тағдыр»** ұғымының» концептісі жатады [7; 9]. Оларға жеке адамның еркінен тыс пайда болған түсініктерге байланысты мынадай терминдер: ***тағдыр, жазмыш, бақыт, үлес, сыбаға, жеребе (жребий), жағдай*** және т.б. жатқызуға болады,

«*Тағдыр*» концептісі өрісінің аясын талдау мақсатында аталған ғалымдар әр түрлі жүйедегі (үндіеуропа және басқа) тілдердің деректері бойынша және түрлі мәдениеті (көне және қазіргі), өз философиясы мен діні бар халықтар тілдерінің материалдары негізінде біршама зерттеулер жүргізді. Әсіресе олардың славян тілдері мен мәдениетіне қатысты зерттеулерін атап өтуге болады. **«*Тағдыр*»** концептісінің өрісі аса кең екені айқын. Егер адамның «өмірлік философиясының» алғашқы көрінісі жекелік және субъективті сипатта болса, екіншісі – объективті, яғни **«шындық»** (ақиқаттық) түсінігіндегі сипатқа ие екені анықталды.

«*Шындық*» пен **«*тағдыр*»** аралығында: ***әрекет*** (действие), ***менталды әрекет*** (ментальное действие) және ***сөйлеу әрекеті*** (речевое действие) тәрізді үш түрлі түсінік орын алады. Ал бұлардың басын біріктіретін **«*әрекет*»** концептісі. **«*Әрекет*»** концептісі саналы қызметпен айналысатын адам баласының өмір сүру әлемінің негізі екені анық. Егер **«*тағдыр*»** адам өмірін алдын-ала анықтайтын болса, **«*әрекет*»** адамның өмір сүруіне жағдай жасайды. **«*Тағдырды*»** адам өзі үшін таңдай алмайды, ал **«*әрекет*»** онымеи салыстырғанда адамның алға қойған мақсатына сай таңдалып алынады. Егер **«*тағдыр*»** адам баласын өмірдің негізгі тұтқасы болуынан ажыратуға тырысса, ал **«*әрекет*»** тіл арқылы жүзеге асып, дүниенің тетігі – **тіл** екендігін білдіреді [10-13].

Адам **«*әрекет*»** арқылы ғана **«*шындық*»** болмыспен қарым-қатынас жасай алады. Сондықтан **«*әрекет*»** – адам мен әлем арасындағы қатынастарды реттеу орталығы болып саналады. Әлем туралы сөз еткенде «**шын**» деген анықтауышты қосып, оны «**шындық болмыс**» деп айтуымыз да бекер емес. **«*Әрекет****тен*»** сөйлеу ісіне (актісіне) өту әрі табиғи, әрі оңай түрде іске асады. Кейбір сөйлеу актілерін **«*әрекет****пен*»** тікелей сәйкестендіру 1960 жылдың басындағы Дж. Остин негізін салған перформативтер теориясынан бастама алады. «Перформатив» сөзі латынның «performo» сөзінен шыққан – ***әрекеттенем, істеймін*** деген түсінік береді. Аталған теория бойынша сөйлеу актісі мына мағынадағы сөздермен теңеседі: ***ант, уәде, серт, жаза, атақ беру, салт, рәсім (салтанат)*** Сөйлеу актісінің құрылымының өзі-ақ, негізінен алғанда, **«*әрекеттің*»** үлгісін (моделін) қайта тудырады. Себебі сөйлеу актісінде әрекеттенушінің ниеті, мақсаты, **«*әрекет*»** нәтижесінің күші мен әсері көрініс табады. Мысалы, өмірде кездесетін саясат, дипломатия, басқару мен юриспруденция (заң ғылымы) және сот ісін

жүргізу және т.б. қызмет салаларына қатысты ***«әрекеттің»*** жазбаша және ауызша түрлерін айырып алу қиындыққа соғады. Айталық, ***бітім*** *жасау* ***(бітім шарт), бұйрық беру, жарлық шығару, наразылық (қарсылық) білдіру*** және т.б. түрлері. Сөйлеудің екі түріне де тән ортақ қасиеттер бар: ***ниет, себеп, дәлел, мақсат.*** Олар ашық не жасырын түрде, қосымша әсерлі, тікелей не жанама нәтижелі, салдар мен бағалау (практикалық және этикалық) түрінде кездеседі. Мәселен, егер дәстүрлі тәртіп нормасы бұзылса, ол сөйлеу әрекеті ме, әлде сөзге қатыссыз істелген әрекет пе – екеуіне де жауапты болатын – адам. Мұндай екі түрлі әрекет жағдайында да адамға бірдей айып тағылуы не тағылмауы (ақталуы) мүмкін. Сөзге қатысты не қатыссыз болуына қарамай, әрекеттер сызықтық (линейность) сипаттағы уақытқа байланысты өрістейді, екеуінде де басы мен соңы және аяқталу сипаты болады. Бұларға ортақ жайт – іс-әрекеттің орындалу стратегиясы. Яғни сөздің іске және, керісінше, істің сөзге айналуы.

Сөз бен істің және сөйлеу актісі мен іс-әрекеттің ұқсас жақтарына қарамай, сөйлеу әрекеті айырықша ерекшелінеді. Оның негізгі айырымдық белгісі – екінші біреуге бағытталу сипатында. Бұл жердегі «екінші біреу» дегеніміз – жеке немесе әлеуметтік қабылдаушы, таныс не жат, дос не қас, замандас не ұрпақ, аруақ, құдай не әулие тіпті әрекеттенушінің өзі де болуы мүмкін. Егер айтылған сөз абсолютті бос кеңістікке бағытталса, ол сөйлеу әрекетіне жатпайды. Ал халыққа не ұлтқа тән этикет пен салт адам баласының мінез-құлқын анықтайды деп саналады.

Сөйлеу әрекетінің басқа әрекет түрлерінен, мысалы, **сөзсіз әрекеттен** мәнді бір айырымдық белгісі – **пікірдің** қатыстылығы, яғни толық не қысқа айтылған пропозиционалды мазмұндағы пікірдің қатысты болуы. Осының нәтижесінде **сөйлеу әрекеті** сөйлеу актісін мезгіл категориясынан алыстататын **менталды әрекетке** барынша жақындайды. Себебі **пікір** уақытқа қатыссыз деп саналады (атемпоралды). Пропозиционалдық мазмұнның қатысуымен **сөйлеу әрекеті** этикалық және утилитарлы бағалаулармен бірге ақиқаттық бағаға да ие болады.

Сонымен, **сөйлеу актілері,** бір жағынан, **сөзсіз (бейвербалды) әрекетпен,** екінші жағынан, **менталды әрекетпен** ортақ белгіге ие. **Сөзге қатыссыз** әрекеттермен оларды жақындататын ең алдымен **мақсаттылық** (целенаправленность). Ал **менталды әрекетпен** жақындататын – **пропозицноналды мазмұн. Сөйлеу әрекеті** – менталды қызмет пен адам қызметі кезінде аралық рөл атқарады да өзара біртұтас кешен құрайды.

Ресейлік зерттеу тобының төртінші бағыты ***«шындық»*** пен ***«шыншылдық»*** проблемаларының кешені [15]. ***«Шындық»*** ұғымы біркелкі емес. ***Діни шындық*** – жер бетіндегі ақиқаттықты (реальность) «ана дүние жаратушысының» ***«шындығымен»*** салыстырудан туындайды. Басқаша айтсақ, ***«шындық»*** ойдың мәні (сущность идеи) мен құбылысқа қарама-қарсы түсінік. ***«Логикалық шындық»*** жалған айтылымға қарсы қойылумен бірге, болмыстық жағдайды менталды категория арқылы пайымдау негізінде анықталады. Қалай болғанда да, ***«шындық»*** уақыттан тыс тұрақты сипатта болады және жалғыздық пен арман әлеміне қатысты белгілер оған аса бір тән

қасиет деуге болады. Сонымен, «***шындық***» – жалғыз, бірақ оны **ақиқат әлемі** және **арман әлемі** деп екі жақты сипатта қабылдаған жөн. ***Арман әлемі ақиқат әлемінен*** туындайды, сол себепті оны екінші кезекте карастыру қажет. ***Шындыктың*** діни және философиялық ұғымдарына қарағанда, ақиқат және жалған айтылымдарды біріне бірін қарама-қарсы қойып қарастыру оларды танымдық нысан, сөйлеу субъектісі деп санау үшін аса ыңғайлы. Сөйлеу субъектісі хабар жеткізуші ретінде өз деректерін әрдайым толықтыра түсуге ынталанудың негізінде шындықты суреттейтін пайымдаулар жасауға талаптанады. Себебі табиғи тіл өз болмысында екі жақты логикамен – ***шындық*** және ***жалған деректермен*** үнемі «күресуде» болады, яғни заңды құбылысқа күмән туғызу мен объективті пікірді субъективтімен ауыстыру жағын көздейді. Сөйтіп, логика ілімі табиғи тілге бір жағынан «күрес жарияласа», екіншіден сол тілдік деректер негізінде «***шындыққа***» жетуді көздейді. Табиғи тілдің «***шындықты***» бұрмалауға қажетті мүмкіндіктері мол деп саналады. Мысалы, модаль сөздерден басқа, оларға қатыстылар: шамалау белгілері ***(азды-көпті, көбінесе, айырықша*** т.б.), жалпылау, топтау ***(жалпы айтқанда, қысқарта айтқанда),*** модальді белгісіздік белгілері ***(сол сияқты, сол тәрізді, мүмкін, ұқсас, дәл, шығар),*** сандық белгілер ***(шамалас, жақын, жуық)*** және т. б. [15].

ТЛТ тобының тағы бір (бесінші) зерттеу бағыты – «***Тіл және уақыт***» мәселесіне қатысты [18]. Бұл бағыт тілдің грамматикалық жүйесіндегі лексикалық көрінісін «***уақыт***» концептісі ретінде қарастырып, айтылым құрылымының сызықтық (линейность) белгідегі «***уақыттың***» бірбағыттық қозғалысын, бірөлшемдік сипатын ескере зерттейді, Ф. де Соссюрше, тіл механизмінде сөйлеудің сызықтық белгісі тілдік белгінің еркіндік сипатымен қатар іргелі принципі деп саналады. Бұл тұжырымдар тілдің ішкі құрылымындағы «***уақытқа***» қойылатын шектеулерді жеңудің аса бір күшті және әсерлі жолы деуге болады.

«**Тіл**» қызметіндегі ішкі байланысты әлсіретуге қатысты көмекші сөздер, есімдіктер, анафоралар және т.б. бірліктер тілімізде жеткілікті. Осы бірліктердің негізінде пікір ақиқаттығына әсер ететін айтылым тұтастағына қатысты «***уақыт***» факторының өлшемі азаяды. Бұл жайт тілдегі байланыстың алшақтық сипаты тек ойлау процесіне ғана емес, сөйлеуге де тән қасиет деп қорытындылауға мүмкіндік береді.

Логика ілімі «***уақыт***» түсінігін «жоққа шығаруға» әрекеттенумен бірге, оны өз зерттеу нысанына алғаны мәлім. Мысалы, «**уақыттық логикалар тілі**» мәселесін модальді логика деңгейінде зерттеу ісі арнайы күн тәртібіне қойылды. Мәселенің ойдағыдай шешілуі ғалам бейнесінің динамикалық құрылымын формалдау проблемасына өз үлесін қосатыны сөзсіз. Осы істің бастамасы ретінде «**әрекет логикасы**» мен «**болжау логикасын**» зерттеу мәселелері қолға алынуда. Мұндай зерттеудің болашақтағы нәтижелері «**уақиға ықтималдығын бағалауға**» да аса бір қажетті деректер беруі мүмкін [16].

Аталған бағытқа қатысы бар тағы бір проблемалық мәселе – тілдің физикалық, әлеуметтік және менталдық кеңістіктегі **«қозғалысының»**

лексикалық және грамматикалық концептуалдану әдістері. Мұндай зерттеулер қазіргі және көне тілдер материалдары негізінде жүргізілуде. Арнайы көңіл бөлінген мәселе – коммуникациядағы символдық мағыналар **«козғалысы»: *ырым, жора, әдет-ғұрып, салт, рәсім, салтанат*** және ұлттық мәдениет пен көркемдік әлемінде орын алатын **«қозғалыстың»** символдық мағыналары.

ТЛТ тобының проблемалық зерттеу мәселелерінің тағы бір (алтыншы) бағыты – **«*уақыт*»** пен **«*қозғалысқа*»** тақырыптық жағынан жақын тұратын **«*кеңістік*»** концептісі [19]. Зерттеу барысында **«*кеңістік*»** пен **«*уақыт*»** аралығындағы өзара қарсы тұлғадағы мынадай проблемалық қатынастар қарастырылды: **«*уақыт*»** – қозғалыста; **«*кеңістік*»** – тұрақты; **«*уақыт*»** – бірөлшемді; **«*кеңістік*»** – үшөлшемді.

«*Уақыт*» пен **«*кеңістікті*»** адам баласы ақиқат шындықтағы материя деп қабылдайтыны мәлім. Осы тұрғыда пайымдасақ: **«*кеңістік*»** концептісі **«*уақыт*»** концептісіне қарағанда көрнекі түрде беріледі, сондықтан **«*кеңістік*»** семантикасы **«*уақыт*»** семантикасына қарағанда бірінші кезекте қабылданады.

Әлеуметтік және рулық қатынастарға, адамның ішкі дүниесіне, оның жекебас әрекеттің өрісіне, этикалық мінез-құлқы, мифтік әлемі, ғылыми білім дәрежесін үлгілеу (модельдеу) процесіне қатысы бар сөздерді әр түрлі топтарға бөліп қарауға болады. Мәселен, заттың кеңістіктегі орналасу қалпын (күйін) және өлшемін білдіретін сөздер ***(биік, аласа, жалпақ, жіңішке, ұзын, қысқа, тік, қисық* және т.б.)**; тұрпатын ***(дөңгелек, ұзынша, квадрат пішінді, куб пішінді*** және т.б.) және басқа да **«*кеңістік*»** сипатын анықтайтын сөздерді атап өтуге болады. Бұлар көптеген метафоралық мағынадағы сөздер жиынының қорын құрайды. Осындай сөздер қорының ішінде «**жол**» сөзінің метафорасы адамның рухани өмірі мен оның мақсатты әрекетін түсінуде аса маңызды рөл атқарады. Бұл сөздер жиыны **зат** пен **кеңістік** әлемі ішінде адам баласының бейімделу мүмкіндігіне қажетті үлгілерді ***(сол, оң, алды, арты, жоғары, төмен),*** ұғымдар мен категорияларды тануда үлкен рөл атқарады. Осы бағыттағы зерттеулерде түрлі тілдерге қатысты кеңістік параметрлерінің семантикасына және олардың ауыспалы мағыналарына көп көңіл бөлінген.

Жоғарыда сөз болған ғылыми топ айналысқан тағы бір бағыт – **«Мәдениет пен тілдегі адам бейнесіне»** қатысты кешенді мәселелер [17]. Егер «жаратушы» (құдай) адам баласын жаратса, адамның өмірге әкелген ең бір зор үлесі – «**Тіл**». Сол сиякты, егер аллатағала өз бейнесін адам кейіпінде қалдыра білсе, адам өз бейнесін **«Тілде»** көрсете білді. Адам өзі не басқа адам жайлы білгенін-жиғанын «**Тіл**» арқылы сыртқа көрсетеді деуге болады. Мәселен, адам өзінің физикалық кейіпі мен ішкі сезімін, қайғысы мен қуаныш сезімін, жанды не жансыз әлемге қатынасын және т.б. тіл арқылы жеткізе білді. Сонымен бірге адам өз ойының бастамасы мен шығармашылық қабілетін түгелдей **«Тілге»** арнады. Сондықтан да адам феноменін ұғыну жолы қоғамдық ғылымдар саласы арқылы ғана емес, оның жолы тек табиғи білім берудің нәтижесінде адам санасы қалыптасады. Сондықтан да XX ғ.

философиялық зерттеу бағыттарының «**Тіл туы**» астында дамуы кездейсоқтық емес. Әрине, бұлайша тұжырымдарға негіз болған жайт: философиялық ойдың нысаны ретінде тек **«Адамның»** алынуы.

Жан, рух, жүрек, ар, ұят, ақыл, ес және т.б. **концептілер** адам затының **«Тілдік бейнесін»** зерттеу нысаны етіп алуға себепші болғаны белгілі. Бұл концептілер әртүрлі мәдениет ареалдарына: орыс халкының мәдениетіне, антик әлеміне, батыс еуропаға, солтүстік халықтарына, қиыр шығыс елдері мәдениеттеріне және т.б. қатыстыра қарастырылды.

«Адам баласының тілдік бейнесі» атты кешенді проблеманың жалғасы ретінде **«Этика тілі»** мәселесі де жеке зерттеу нысаны болғанын айта кету қажет [20]. Зерттеу барысында мынадай этикалық концептілерге көп көңіл бөлінді: ***жақсылық, зиянкестік, әділеттік, ұят, ар, парыз, күнаһарлық, масқаралық (қара беттік), ақау, кеселдік, ізгілік, тазалық*** және т.б. Аталған концептілер еуропа және шығыс елдері мәдениетіне қатысты әлі де зерттеле түсуі керек. Осымен қатар діни әдебиет тіліндегі, рухани әлемдердегі, көркем әдебиеттегі және күнделікті сөйлеу тіліндегі аса құнды пікірлер де арнайы зерттеліп жатыр [20].

ТЛТ тобының сегізінші проблемалық зерттеу нысаны **«аяғы»** және **«басы»** концептілерінің семантикасы [21]. **«Аяғы»** мен **«басы»** сияқты **«ескі» мен «жаңа», «бірінші»** мен **«соңғы», «бастауы»** мен **«ақыры»** ұғымдары концептуалды өрістердің семантикасы тұрғысынан ерекше назар аударуды қажет ететін зерттеу нысаны болып табылады. Әсіресе **«аяғы»** мен **«басы»** ұғымдары соңғы кезең зерттеулерінде жиі қаралды. Бірақ бұл екі ұғым жансыз табиғи әлемінен гөрі жанды әлемге тән деп саналады. Мысалы, өсіп тұрған ағашқа, таудың және т.б. табиғи ортаның **«аяғына»** байланысты: ***ағаштың аяғы, үйдің басы (аяғы), таудың аяғы*** деп қолданыла бермейді. Ал ***өзеннің бастауы, өзеннің сағасы, таудың етегі, таудың төбесі*** және т.б. сияқты қолданыстарды көптеп кездестіруге болады. Табиғатты құраушылар үшін адам түсінігінде **«аяғы»** мен **«басы»** концептілері қарама-қарсы қойылмай, олар **«бүтін», «бөлек»** терминдерінің мағынасымен беріледі. Мысалы, ***саусақтың ұшы, мұрынның ұшы*** деп адамның дене мүшесінің бөлігін атайды, ал бұл жерде шекаралық мағына жоқ деуге болады. Тек Эвклид пен Платонның әлемнің геометриялық моделдері мен уақыт қозғалысының сызықтық және бір бағыттық сипаты ғана **«аяғы»** мен **«басы»** ұғымдарының қалыптасуына негіз бола алды. Бұл түсініктер мезгілдік процестегі уақыт ағымына да сызықтық пен бағыттық белгілері бар **«кеңістік»** нысандарына ***(жолға, көшеге, соқпақ жолға, сүрлеу жолға*** және т.б.) бірдей таралады. Аталған ғылыми бағыттағы негізгі және кешенді мәселелеріне: етістіктің түріне қатысты проблемалар (проблемы вида глагола) және созылыңқы әрекетті редукциялап, оны уақыт осіндегі бір нүктеге беттестіру мүмкіндігі, **«басы»** мен **«аяғы»** концептілерінің адам әрекеті кезінде саналы немесе өздігінен пайда болатынын анықтау мәселелері, сонымен қатар, **«басы»** мен **«аяғы»** араларындағы «ескіру» мен «жаңғырту» семантикалары және басқа да көптеген коннотация мәселелері жатады. Мысалы, салыстырыңыз; бір істі ***бітірудің – аяғы (соңы)*** және ол

істі ***қиратудың (жоюдың) – аяғы (соңы), мақсатқа жетудің (жетпеудің)– аяғы (соңы), ұтудың (ұтылудың) – аяғы (соңы)*** және т.б.

ТЛТ тобының 2000 жылдары кең өріс алған зерттеу бағыты **«космос»** және «**хаос**» проблемалық тақырыбы **«реттілік»** (порядок) және **«ретсіздік»** (беспорядок) концелтуалды өрістерін қарастыруға арналды [22]. Мұндай екі қарама-қарсы ұғымдар өрісінің біреуі ғалами «**космос**» **(ғарыш)** концептісін ***(космос/ғарыш, реттік, тәртіптілік, норма, заң, заңдылық, үйлесімдік, ұйымдастырушылық, ұқыптылық*** және т.б.), ал екіншісі **«хаос»** концептісін ***(хаос, ретсіздік, тәртіпсіздік, аномалия, ауытқу, тәртіпті бұзушылық, жөнсіздік, кездейсоқтық, үйлеспеушілік*** және т.б.) құрайды.

Аталған ғылыми топтың оныншы реттегі проблемалық зерттеу мәселесі **«сұлулық»** концептісі және оның негізгі ұғымы **- «сұлулық өрісі».** Ол **шындық, жақсылық, сұлулық** концептілерінің кешенді үш тағанының басын біріктіретін әр тілдердегі мәдени эстетикалық мағыналарының көріністерін қарастырады.

Зерттеу мақсатын жалпылама тұрғыдан қарағанда, мұнда лексикалық, сиитаксистік, интонациялық бағалауды (оң, теріс) білдіруші мүмкіндіктердің талдауы мен сипаттаулары да кездеседі.

Жоғарыда айтылғандарды қорыта келе, ТЛТ ғылыми-теориялық зерттеу тобының тілді логикалық талдау бағыты бойынша ерекше көңіл бөліп, болашақ ізденістерде жалғастыруды көздеген мәселелері мынадай:

1) түрлі мәдениет пен өнер теориясындағы **«сұлулық»** концептісінің дефициясына (пайымдауларына) назар аудару;

2) өнер түрлерінің (шешендік, бейнелеу өнері, музыка саласы) шындық нысандары мен көркемдік бейнелеулеріндегі эстетикалық баға берудің түрлерін қарастыру;

3) табиғи және рухани әлемді эстетикалық тұрғыдан бағалау және эстетикалық бағалаудың шегін анықтау;

4) сымбаттылықтың статика мен динамикаға, хаос пен реттілікке қатысын арнайы қарастыру;

5) табиғаттың жанды және жансыз нысандарының әр типтеріне эстетикалық баға беретін тілдегі метафоралар мен басқа да бейнелеу мүмкіндіктерін қарастыру;

6) эстетикалық бағаны білдіретін пайымдаудың мүмкін болатын түрлерін (верификации) зерттеу;

7) өмір мен өнердегі эстетикалық бағаның диахронды өзгерулеріне назар аудару; өнертану, көркем әдебиет және публицистикаға қатысты қазіргі мәтіндер, қоғамдық (топтық) ауызша сөйлеу үлгісі, диалектілік және фольклорлық, этимологиялық материалдар, тарихи ескерткіштер мен көне тілдердің деректері осы зерттеулерге негіз болатын тілдік материал ретінде алынды.

Тілді логикалық тұрғыдан зерттеуге қатысты ғылыми шолудың тағы бір мақсаты – аталған кешенді мәселелерге қазақ тілін зерттеуші ғалымдардың назарын аударып қана қоймай, сонымен бірге қазақ тіл білімінің зерттеу аясын кеңейте түсуге және тілді тану мәселесінде тілдік

деректерді өзіміз өмір сүретін ақиқат шындықпен тығыз байланыста қарауға сеп болу. Әрине мұндай мақсатқа жету үшін тілдік жүйенің құрылымдық бірліктерін философия, психология, этностану, әлеуметтану, мәдениеттану, елтану, антропология сияқты ғылым салаларымен байланыста қарау қажет болады.

Бұл ғылыми сала қазақ тілін зерттеуші ғалымдардың назарынан толығымен тыс қалып келеді деген ұғым тумауы керек. Мәселен, тілші-ғалым А.Исламның қазақ тіліндегі «Өмір», «Бақ» концетілерінің көрінісі жайлы ізденістері тілшілердің назарын аударарлықтай. Осы мәселе жөнінен М.Т.Күштаеваның «Тары» концептісінің семантикалық құрылымы мен лингвомәдени мазмұны» атты кандидаттық диссертациясын атауға болады. Зерттеуде қазақ тіліндегі тары дақылына қатысты тілдік бірліктердің семантикалық құрылымы және осы тары атауына қатысты тұрақты тіркестердің концептілік өрісі қарастырылған. Бірақ біз келісе қоймайтын жайт ол жұмыс тақырыбындағы «**Тары концептісі**...» сөз тіркесінің кездесуі. Қайшылық зерттеушінің өзі негіз етіп алған (сүйенген) «концепт» ұғымының «Адамның әлем туралы жинақталған мәдени түсініктері бейнеленген атауларды бар ғаламның тілдік бейнесі ретінде тану» мазмұнындағы анықтамасынан көрініп тұр дер едік. «Тары» атауы ұрпақтан-ұрпаққа тарауды айтпағанды, қазақ елінің солтүстік өңірінде тіпті де қолданылмайды деуге болады. Сол сияқты, мысалы, қой шаруашылығына байланысты ұлттық қолданыстарды – «қой концептісі», өзбек халқының мақта шаруашылығына қатысты – «мақта концептісі», орыс ұлтына қатысты – «картөшке концептісі», америка халқының жүгері шаруашылығына сай – «жүгері концептісі» және т.т. атала беру дұрыс болар ма еді? Әрине, аталған ғылыми жұмысты талқылау бұл мақаланың міндетіне кірмейді.

Сонымен, қазақ тілін логикалық тұрғыда талдауды жүйелі сипатқа айналдыру заман талабы демекпіз.

Әдебиет

1. Пропозициональные предикаты в лингвистическом и логическом аспекте. Тезисы докладов конференции. М., 1987.
2. Прагматика и проблемы интенсиональности М.,1988.
3. Референция и проблемы текстообразования М.,1988.
4. Логический анализ языка: Знание и мнение. М., 1988.
5. Логический анализ языка: Проблемы интенсиональных и прагматических контекстов М., 1989
6. Логаческий анализ языка: Противоречивость и аномальность текста. М., 1990.
7. Концептуальный анализ: методы, результаты, перспективы. Тезисы докладов конференции. М.,1990.
8. Тождество и подобие, сравнение и идентификация. М., 1990.
9. Логический анализ языка: Культурные концепты. М., 1991.

10. Действие: Логические и лингвистические модели. Тезисы докладов конференции. М., 1991.
11. Логический анализ языка: Модели действия М., 1992.
12. Логический анализ языка: Ментальные действия. М., 1993.
13. Логический анализ языка: Язык речевых действий. М., 1994.
14. Понятие судьбы в контексте разных культур. М., 1994.
15. Логический анализ языка: Истина и истинность в контексте разных культур. М., 1995
16. Логический анализ языка: Язык и время. М., 1997.
17. Логический анализ языка: Образ человека в культуре и языке. М., 1999.
18. Логический анализ языка: Языки динамического мира. Дубна, 1999.
19. Логический анализ языка: Языки пространств. М., 2000.
20. Логический анализ языка: Языки этики. М., 2000.
21. Логический анализ языка: Семантика начала и конца. М., 2002.
22. Логический анализ языка Хаос и космос. Концептуальные поля порядка и беспорядка. М., 2002.
23. Вежбицкая А. Язык. Культура. Познание. М.: Русские словари, 1996. – 416 с.

ҚАЗАҚ МӘТІНІН АБЗАЦ ДЕҢГЕЙІНДЕ ЗЕРТТЕУДЕГІ МАЗМҰНҒА ҚАТЫСТЫ ТІРЕК СӨЗДЕРДІ СТАТИСТИКАЛЫҚ ЖОЛМЕН АНЫҚТАУ

ҚР БҒМ Актөбе мемлекеттік педагогикалық институты
«Жұбанов тағылымы» IV Республикалық ғылыми-теориялық конференция материалдары. 14-16 қазан, 1999 жыл. Ақтөбе 2004, 99-105-бб.

Профессор Құдайберген Жұбанов өмір сүрген қысқа мерзім ішінде оның филологияға қосқан теориялық және практикалық шарапатты үлестері қазақ тіл білімінің барлық салаларына дерлік игі әсер етті. Қазақ тіл білімінің келешегіне көз салған ғалым, көптеген мәселелерді өзіне тән көрегендікпен, табиғи дарындылықпен дұрыс шешімін тапты. Тіл білімінің соңғы жаңалықтарының біріне жататын сөйлем құрастырушы элементтердің нақтылы бөлінуі туралы қағиданың заңдылығын танып біле білді. Сол отызыншы жылдардың өзінде-ақ Қ.Жұбанов қазақ тілінің грамматикалық құрылысынан аналитизмнің алатын орнын дұрыс бағалай білді. Тіліміздегі күрделі көмекші етістіктердің табиғатын дәлелдей отыра, олардың мағыналық құрылымын саралап, талдап кеткен.

Ғалымның түсіндіруінше, тіліміздегі қарым-қатынасқа түскен сөздердің мағыналық жағын танып қана қою жеткіліксіз, сол мағына қалайша туындап тұр, қандай формалардың әсері болды, яғни мағына мен форманың аралық тығыз қарым-қатынасын ескеріп отыру қажеттігіне жұртшылықтың

назарын аудартады. «Жаңа грамматиканың жаңалықтары жайында» атты еңбегінде ол былай дейді: «Грамматиканың мақсаты жалғыз-ақ сөздің қандай мағыналары болатындығын баяндау ғана емес, қандай формаларға қандай мағыналар сәйкес келетінін баяндап, дұрысы, қандай мағына шығу үшін сөзді қалайша құрастыру керек екендігін баяндау. Сондықтан құрастыру, құрылыс жағын елемейтін грамматика да болмақ емес» [1, 220]. Ғалым «құрастыру, құрылыс» (структура) ұғымдары арқылы грамматиканың үлкен бір саласы синтаксисті осылайша түсіндіреді. «Елуінші жылдардан бастап, грамматикада өзінше бір жеке сала болып бөлініп шыққан сөз тіркесі синтаксисі де Қ.Жұбановтың назарынан тыс қалмады», - деп жазады синтаксис саласының белгілі маманы Қ.Есенов [2, 71]. Осы маңызды мәселеге арнап проф. Қ.Жұбанов «О формах сочетания слов в казахском языке» атты мақала жазғаны да бізге мәлім. Тағы бір көңіл бөлерлік жайт – ол ғалымның сөйлем мүшелерінің орналасу тәртібіндегі заңдылықтарды, сондай-ақ кісі аттары мен біріккен сөздердің табиғаты кеңінен сөз болуы. Қысқаша айтқанда, профессор Қ.Жұбанов қазақ тілінің өзге салаларымен бірге тіліміздің синтаксис құрылысын зерттеуде үлкен теориялық еңбек қалдырып, жаңа бағыттағы болашақ зерттеулерге аса білгірлікпен жол салып кетті.

Біз сөз етпек болып отырған «Мәтін лингвистикасы», «Мәтін теориясы» атты тіл біліміндегі жаңа бағыт үшін де Қ.Жұбановтың жоғарыда айтылған және т.б. зерттеулерінің де зор маңызы бар. Сол сияқты, тіл білімінің белгілі ғалымдары қатарында, Қ.Жұбанов та тілді зерттеуде, тілдік фактілерді баяндауда сандық деректерге, ұғымдарға жүгінудің қажеттігіне көңіл бөлгендігі өз зерттеулерінен байқалып отырады. Қазақ тіл біліміндегі «Лингвистикалық статистика» деп аталып жүрген жаңа саланың алғашқы нышандары мен белгілерінің көрініс тапқандары туралы белгілі ғалым, филология ғылымының докторы, математик Қ.Б.Бектаев былай дейді: «Лингвистикалық статистикаға қатысты элементтерді – сандық деректер мен мәліметтерді, оны пайдалану нәтижесіндегі кейбір пікірлер мен тұжырымдарды біз проф. Қ.Жұбановтың еңбектерінен жиі кездестіреміз. Ғалымның сол кездің өзінде-ақ тіл білімінде сандық деректерді пайдалану қажет екендігіне ерекше ден қойғанын айқын байқаймыз» [3, 108-116].

Қазіргі кезде ғалымдар мәтін құрайтын элементтердің заңдылықтарын ашу жолдарын қарастырып, оны анализдеуден өтіп, синтездеу әдісімен көбірек шұғылданып жүр. Оның басты себебі, компьютерлік лингвистиканың әдістерімен мәтіннің «туындау» заңдылықтарын анықтап, мәтін мазмұнын автоматты түрде ашып, одан аннотация мен реферат алу қажеттігі және т.б. қажеттік етуде.

Мәтіндік деңгейдегі зерттеулер түркі тіл білімінде (оның ішінде қазақ тіл білімінде) әлі де болса өз деңгейінде қолға алынбай жүрген мәселенің бірі. Мәтін грамматикасының кейбір теориялық мәселелері Н.З.Гаджиеваның, К.М.Абдуллаевтың жұмыстарында кездеседі. Түркі тілдерінің синтаксистік құрылымын зерттеуде аналитизмге бас бұру тенденциясы алғашында проф. Қ.Жұбановтың, бертін келе Н.С.Поспеловтың,

Б.А.Серебренниковтың, Н.О.Оралбаеваның және Н.Х.Демесинованың және т.б. ғылыми ізденістерінен кездестіруге болады.

Мәтіндердің құрылу (туындау) заңдылықтарын анықтау, әсіресе олардың ішінен мағына жағынан ең негізгі деген тілдік единицасын айыра білудің қазақ тілін үйрету (үйрену) процесінде де тәжірибелік мәні зор. Ол үшін ең әуелі мәтіндерді сегменттеу ережелерінің көмегімен үлкенді-кішілі мәтіндік единицаларға топтаудың қажеттігі туады. Зерттеушілердің тәжірибесінде мәтіннің құрылымына қатысты единицалар деп: сөздер, сөз тіркестері, сөйлемдер, абзацтар, параграфтар, бөлімдер, тараулар, томдарды атап жүр.

Табиғи өмірде кездесетін кейбір объектілердің арасындағы қарым-қатынастарды айқындап білу барысында, жеке сөздерге қарағанда сөз тіркестерінің мәні ұтымды деуге болады. Объектілердің қарым-катынасын анықтайтын сөз тіркестері мәтін ішінде атрибутты, объективті, предикативті, адвербиалды болып бөлінеді.

Ғалым Қ.Жұбановтың қазақ тіліндегі сөз тіркестеріне аса көңіл бөлуі – сөз тіркестерінің синтаксис саласының пайда болуына үлкен себін тигізді. Оның сөз тіркеске байланысты пікірінің құндылығы, сөздердің өзара байланысуында жарыспалы (параллельное) және тізбекті (последовательное) амалдарын ашып көрсетуінде. Сонымен бірге, сөйлем мүшелерінің бірінші қатардағы (первичное), екінші, үшінші т.б. қатардағы кездесетіндігін байқағандығының да мәні зор. Сөйлем ішіндегі сөздердің біразы бір-бірімен тікелей тіркессе, енді біреулерінің баспалдақтап тіркесу мүмкіндігін атап өтеді. Осындай құнды деректер біздің ізденістерімізге де көмегі тиді деуге болады.

Енді ғалым-тілшілердің кейінгі зерттеулері, қазақ тілінің сөз тіркесінің синтаксисінің пайда болуы тәріздес, қазақ тілінің жаңа саласы – мәтін синтаксисіне әкелуі де күмән туғыза қоймас.

Біздер көбінесе өз ізденістерімізде индоеуропа тілдеріне қатысты зерттеулердегі үлгі-тәсілдерге жүгіндік. Түркі тілдерінің оның ішінде қазақ тіліндегі мәтіннің құрылымын абзац деңгейінде зерттеу бұрын-соңды жүргізіле қоймағанын еске сала кетейік. Жалпы ғалымдардың пікірінде мәтін ішінде абзацтың қандай единицаға жататындығын анықтауда әлі де болса біріздiлік байқалмайды. Кейбір зерттеушілер оны композициялық десе, енді біреулері логикалық немесе семантикалық единица қатарына жатқызып жүр. Біздің түсінуімізше абзац – бір орынға жай ғана топталған сөйлемдердің жиынтығы емес, ол автордың өз ойын оқушының саналы түрде қабылдауына, оны түсінудің амалдарын жеңілдетудің мақсатына лайықтап құрылған сөйлеу тәжірибесінде бар мәтіндік единица.

Абзац мәселесін мазмұн тұрғысынан қарастырсақ, мәтінде қалам шеберінің ойы шағын-шағын үзік хабарламалар арқылы беріледі де, олардың бірігуінен барып қомақты хабар пайда болады. Әр мәтіннің тақырыбы іштей кішігірім тақырыпшаларға бөлініп, қазіргі әдебиеттерде аталып жүрген субтема және микротемаларға ажыратылады. Мәтін тақырыбы белгілі-бір

қатынастық байланыста тұратын бірнеше микротемалардың қосындысы ретінде қаралады.

Ал абзацты құрайтын сөйлемдер бір микротеманың немесе бір-бірімен тығыз байланыста тұрған бірнеше микротеманың мазмұнын анықтайды. Осы абзацтың ішіндегі сөйлемдер бір-бірімен өзара арнаулы лексикалық амал-тәсілдер арқылы және белгілі синтаксистік байланыста болулары шарт.

Айтылу мақсатына сай сөйлемдері: хабарлы, сұраулы, бұйрықты және лепті сазды сөйлемдер деп жіктеп, табиғатын ашу сияқты сөйлеудің (речь), функционалды-мағыналық түрлеріне қарай мәтін абзацтарын да хабарлау, сипаттау, пікір тудыру абзацтары, сұрақ мәнді абзацтар, ниет пен ойды білдіретін абзацтар және аралас абзацтар деп бес топқа бөлуге болады.

Құрылымдық тұрғыдан қарағанда мәтін абзацтарының лексикалық жағынан толыға түсуі, синтаксистік және грамматикалық жағынан түрлене түсуі сол абзацтың мәтін бойындағы орнына көп байланысты. Тұрған орнына қарай абзацтарды кіріспе (бастапқы), ішкі және қорытынды абзацтар деп те үш түрлі қарастыруға болады.

Синтаксистік зерттеулердің сөйлемнің басында күрделіленген түрлермен ғана шектемей, зерттеу шеңберін кеңейтіп, мәтін құрылымына бас бұрып, оның негізгі деген единицаларының және олардың арасындағы күрделі байланыстардың сыр-сипаты мен шарттарын белгілеу өзінше бір проблемаға айналып, мәтін мазмұнына тікелей байланысты оның тақырыбын ашу қажеттігі туды. Осы мәселені шешу барысында мәтінге тән жалпы тақырыптың іштей шағын тақырыпшаларға (подтема, субподтема, микротема) ажырату қолайлы деп есептелінді.

Мен, өз баяндамамда, қазақ мәтінінің мазмұн құрайтын негізгі единицаларын статистикалық жолмен ажырату жағына ғана тоқталмақпын.

Автордың белгілі бір жайды оқырманға тұжырымдап айқын жеткізу кезінде сөздерді орын-орынмен, рет-ретімен таңдап білудің жолдарын анықтау да өзінше – бір проблема.

Сөздер жиынтығын үлкен екі топқа бөліп, олардың біріншісін – мағыналық сөздер тобы, ал екіншісін – мәнерлеуші сөздер тобына жатқызып қарастырғанды жөн көрдік. Мағыналық топтағы сөздердің сөйлемдегі синтаксистік құрылысты қалайша толықтыра түсу жағын (наполняемость) өз алдына зерттей келе, оның абзац пен мәтін көлеміндегі сөздердің жиілігі және сөздердің өзара байланыс күшіне қатысты екенін, ол жағдайдың арнайы жазылған логикалық бағдарламалар арқылы іске асатындығын атап айтқымыз келеді.

Ең бірінші, біздің зерттеуіміздегі сөз болатын «мәтін» ұғымын қалай түсіну керек?

«Мәтін» – біздің анықтамамыз бойынша, іштей бөлшектелінбеген жәй ғана әңгіме, повесть, очерк, поэма, өлеңдер, мәтін бойындағы параграфтар (ішкі бөліксіз), ғылыми және көркем-әдебиеттер, құжаттық шығармалар, сонымен бірге осы тәріздес ірі шығармалардың үзінділері деп ұғынамыз.

Мазмұн жағынан қарастыру барысында мәтін статистикасы және мәтін динамикасы деген ұғымдар ескеріледі. Біз осы баяндамамызда мәтін теориясын тек статика тұрғысынан ғана сөз етпекпіз.

Мәтінді мағыналық құрылым жағынан қарастырған жағдайда, мағыналық бірлік сол мәтін ішіндегі белгілі-бір «тірек сөздер» деп аталатын сөздерге байланысты (орысша әдебиеттерде олар: опорные слова, топики, смысловая веха, дескрипторы, ядра және т.б.).

Негізгі тірек сөздерді (НТС) айыруда мынадай шартқа жүгіну керек сияқты:

1) зерттеп отырған мәтінде ол сөздер жоғары жиілікте болуы қажет;

2) тірек сөздері бар абзацтардың саны (жиілігі) максимум дәрежесінде болу қажет;

3) тірек сөздердің санына контекстік синоним сөздер мен сөздердің есімдіктік орынбасарлары қоса есептелуі қажет;

4) негізінен, мәтінде ұшырасатын сөздердің табиғаты – баяндалатын уақиғаға қатынасатын кейіпкерлердің және сол уақиғаның болған орны мен уақытын білдіретін атаулар болғандықтан, олар зат есім сөздердің (не сөз тіркестерінің) ішінен алынуы қажет.

Негізінен алғанда, уақиға осындай негізгі тірек сөздерімен (НТС) айқындалатын жағдай болғандықтан, олар бұдан төмендеу сатыда тұратын объекті не субъектілермен қарым-қатынаста болатыны сөзсіз, сондықтан олардың атауларын біз көмекші тірек сөздер (КТС) деп атауды жөн көрдік.

Осы айтылған НТС және КТС мәтіннің негізгі статикалық мазмұны (НСМ) деп аталатын кестеде көрсетілді, ал мәтін бойында бар, бірақ НСМ кестесіне енбеген сөздер жиынтығын – қалған сөздер (ҚС), не басқа сөздер (БС) тобына жаткыздық. Біз тек осы кестенің жасалу жағына ғана тоқталайық. Мәтін мазмұнын анықтайтын «маңызды дәрежедегі» сөздер мен сөз тіркестерін анықтаудың бірнеше жолдары бар. Солардың ішінен ең ыңғайлы деп біз терминнің статистикалық тұрақтылық коэффициентін есептейтін формуланың өзгертілген түрін пайдаландық [4, 82-89]:

$$K_{\text{важ}} = \frac{\text{Fm}}{\text{Nn}} \ . \qquad (1)$$

осы формуладағы: F – мәтіндегі сөздердің абсолютті жиілігі (синоним сөздер мен есімдік сөздерді қосып есептегенде), m – алынған сөзге қатысты абзац саны; N – мәтіндегі сөз қолданыстың жалпы саны; n – мәтіндегі абзацтың жалпы саны.

Негізгі тірек сөздер (НТС) мен көмекші тірек сөздердің (КТС) шекарасын айқындау үшін төменде көрсетілген тәжірибе жүзінде анықталған теңсіздіктерді пайдаландық (2), (3);

$$\frac{\left(\frac{n}{m_1}+1\right)^2}{N \cdot n} \le \text{K}^1\,\text{важ} \le 1, \ (2); \quad \frac{\left(\frac{n}{m_2}+1\right)^2}{N \cdot n} \le \text{K}^2\,\text{важ} < \frac{\left(\frac{n}{m_1}+1\right)^2}{N \cdot n}, \qquad (3).$$

мұнда (2) – теңсіздік арқылы НТС шекарасы анықталады да, ал (3) – теңсіздік арқылы КТС шекарасын есептеп шығаруға болады.

Бұл формулалардағы m_1 және m_2 қазақ мәтіндері үшін тәжірибе үстінде анықталатын коэффициенттер. Олардың шамасы мәтіндегі абзацтың жалпы санынан байланысты құбылып отырады. Мысалы, $K^1_{важ}$ үшін мәтіндегі абзацтың саны:

100-ден асса m_1= 42,

81 мен 100 аралығында m_1= 30,

51 мен 80 аралығында m_1= 16,

31 мен 50 аралығында m_1= 17,

15 пен 30 аралығында m_1= 12,

9 бен 14 аралығында m_1= 5,

6 мен 14 аралығында m_1= 4 және т.б.

Осы сияқты тәжірибе жүзінде m_2 шамалары (көрсетілген сегменттер бойынша) басқаша мәндерді қабылдайды. Біздің ұйғаруымызша, бұл коэффициенттердің шамасы тілдік ерекшелікті анықтау критерийі болуы да мүмкін.

Біздің зерттеуіміз қазақ мәтінінің үш түрлі стильдері (публицистикалық, көркем-әдебиет және көпшілікке арналған ғылыми-техникалық) бойынша жүргізілді.

Публицистикалық стиль ретінде «Қазақстан мұғалімі» газетіндегі 1986 жылдың қаңтар айы мен 1987 жылдың қаңтар айы аралығындағы мектеп реформасына байланысты: «Совет мектебі жаңа кезеңде», «Алты жасарлар мектепке барады» атты рубрикамен басылып шыққан 35 мақала қарастырылды. Мақаланың авторлары Қазақстандағы мектеп мұғалімдері, арнаулы газет тілшілері, мұғалімдер білімін жетілдіру институттарының қызметкерлері және тілші ғалымдар. Қарастырылған 35 мақаланың мәтіні 503 абзацтан тұрады.

Көркем Әдебиеттерге жататын мәтіндер қазақтың белгілі жазушыларының шығармаларынан алынған келесі әңгімелері:

1) С.Сейфуллин. «Бандыны қуған Хамит», 102 абзац;

2) Б.Майлин. «Қанды кек», 98 абзац;

3) М.Әуезов. «Көк серек», 97 абзац;

4) С.Мұқанов. «Әсия», 47 абзац;

5) Ғ.Мүсірепов. «Жапон балладасы», 92 абзац;

6) С.Ерубаев. «Бахыт», 53 абзац.

Осы аталған алты әңгімеден барлығы 500 абзац құрайтын 20 шағын мәтіндер зерттелді.

Ғылыми-техникалық стильге жататын мәтін ретінде көпшілік оқырмандарға арналған екі монографиялық шығарма қарастырылды [5]. Бұл шығармалардан алынған мәтін ұзындығы да 500 абзац.

Сонымен үш стильге қатысты мәтіндер бойынша 1500 абзац құрылымы зерттелді.

Тәжірибенің алғашқы қадамдары төмендегі жұмыстарды қамтиды:

1. Аталған мәтіндердегі сөз қолданыстарды сөз табына қатысты белгілеу.

2. Сөзтұлғаларды реестрлік лексема түріне келтіріп сегменттеу.

3. Әр мәтіндегі мынадай төрт түрлі синоним сөздерді айырып белгілеу: лексикалық синонимдер, контекстік синонимдер, есімдік синонимдер, ассоциативті синонимдер.

4. Әрбір бөлек стильдер бойынша абзацтардың алфавитті-жиілік сөздіктерін түзу.

5. Стильдер бойынша әр тақырыпшаларға қатысты сөздердің жалпы жиілігі мен абзацтар санына тиісті жиілігі бөлініп көрсетілетін – үлестірімді алфавитті-жиілік сөздікті түзу.

6. Үлестірімді сөздік бойынша сөздер жиілігінің кему ретімен орналастырылған жиілікті сөздігін түзу.

7. Негізгі сөздердің потенциалды сөздігін түзу.

8. НТС, КТС тізімдері бойынша мәтіндердің негізгі статикалық мазмұн кестесін құру.

Сонымен, бұл ізденіс қазақ мәтіндерінің негізгі сипаттамаларын дұрыс ашу үшін және әрі қарайғы абзац шеңберлеріндегі семантикалық және синтаксистік зерттеулердің бірден-бір бастамасы болмақ.

Әдебиеттер:

1. *Жубанов Х.К.* Исследование по казахскому языку. – Алма-Ата, 1966.

2. *Есенов Қ.* Профессор Қ.Жұбанов және қазақ тілінің синтаксисі // Құдайберген Жұбанов және қазақ совет тіл білімі. – Алматы, 1990.

3. *Бектаев Қ., Белботаев А., Смаилов Т.* Қазақ тіл біліміндегі лингвистикалық статистиканың алғашқы көріністері // Құдайберген Жұбанов және қазақ совет тіл білімі. – Алматы, 1990.

4. *Марусенко М.А.* О формировании словника словаря статистически устойчивых научно-технических терминов // Структурная и прикладная лингвистика. – Вып. 2. – Л., 1983.

5. *Қалыбеков Т.* Экология және кен. – Алматы, 1988; Көшербаев К. Минералдар әлемінде. – Алматы, 1986.

ҚАЗАҚ МӘТІНІНІҢ ЫҚТИМАЛДЫ-СТАТИСТИКАЛЫҚ МОДЕЛІН ҚҰРУ МАҢЫЗДЫ МӘСЕЛЕ

«Академик Ә.Т. Қайдар және тіл білімінің мәселелері» атты халықаралық ғылыми-теориялық конференция . – Алматы, «Дайк-Пресс», 2004 ж., 579-586 бб.

Қолданбалы лингвистиканың көптеген маңызды мәселелерінің бірі – мәтін ішіндегі тілдік бірліктердің орналасу моделінің ықтималды-статистикалық әмбебап сызбасын құру. Мұндай күрделі істі автоматты түрге келтіру: ақпаратты іздеу, тілдерді тиімді жолмен үйрету, базалық (негізгі) тілді құру, тілдің «нормалық» табиғатын түсіну, тілді стилистика тұрғысынан

зерттеуде және т.б. кешенді мәселерді шешуде аса қажетті деп саналады[1].

Тілді математикалық модельдеу кезінде тәжірибе жүзінде алынған сандық деректерге статистикалық талдау жұмысын жүргізу болып саналады. Мұндай талдау арқылы кейбір тікелей байқауға жатпайтын бірқатар тілдік жайттардың ара-жігі анықталады. Ал тілдің ықтималдық моделі көмегімен мәтін ішіндегі кейбір бірліктердің сандық қасиеттерін сипаттауға мүмкіндік туады. Егер осы анықталған бірліктер сипаттамасын ақиқат шындықтың көрсеткіші ретінде қабылдасақ, онда модельден нақты нысанға өту іске асады. Ол үшін зерттеуге алынған мәтін бірліктерінің сандық параметрлерін анықтай білу қажет. Бұл параметрлерге «арифметикалық орта» ($\overline{X}$) мен «стандартты» (S) жатқызуға болады. Әрі қарай $\overline{X}$ пен S мәндері ықтималдық модельдің осы тәріздес параметрлерімен салыстырылады. Тілдік бірліктердің сипаттамалары зерттеушіні сол тәжірибеге түскен шекті көлемдегі нақты мәтінге қатысты ғана қызықтырмайды, ол бас мәтін жиынтығы сипаттамасы тұрғысынан қарастырғанды қалайды. Бұл жердегі «бас мәтін жиынтығы» (генеральная собокупность текстов) дегеніміз тілдің толық түрі немесе бөлек бір стиль, тіл қабаттары, жазушы тілі және т.б. Бұлайша қарастырудың себебі аталған «бас мәтін жиынтығын» қамтитын толық түрдегі мәтін бойынша статистикалық талдау жүргізу тәжірибе жүзінде мүмкін бола бермейді. Яғни зерттеуге жататын нысан «бас мәтін жиынтығының» таңдалып алынған жеке бөліктері (частные выборки) немесе «таңдама бөліктер». Міне, осындай таңдама мәтін бөліктер параметрлерін бас мәтін жиынына қатысты статистика-математикалық зерттеу негізінде ықтималды модельден тілдік нысанның өзіне көшу (өту) жүзеге асады. Мұндай «өту» процесі нақты (белгілі) критерийлерге сүйеніп, статистика-лингвистикалық болжамды тексеру (сынау) негізінде бірліктердің мәтін ішінде кездейсоқ оқиға ретінде «үлестірілу» (таралу) тұрпатын анықтау арқылы іске асады.

Мәселен, мынадай статистика-лингвистикалық болжамды сынау қажет болды делік: мәтін ішіндегі зат есімге қатысты сөздер математикалық статистиканың «нормальды үлестірілу» деп аталатын теориялық заңдылығына бағынады. Мұндай болжамды тексеру үшін зат есім сөздердің эмпирикалық (тәжірибелік) үлестірілу деректерінің теориялық үлестірілу заңдылығы – «нормальды» заңдылығымен сәйкестік дәрежесін тексеру қажет. Ал ондай тексеру тек қана зат есім сөздердің таңдама бөліктер ішінде қайталану жиілігінің орта шамасының бірнеше «сериялары» (мүмкіндіктері) бойынша қарастырылады.

Кездейсоқтық жағдайдағы лингвистикалық шаманың мәтін бойында таралу (үлестірілу) тұрпатының не параметрлерінің сынауға ұсынылып отырған болжамға сәйкестігі нөлдік (негізгі) болжам (H_o) немесе оған қарама-қарсы қойылатын альтернативті болжам (H_1) ретінде тұжырымдалуы мүмкін. Негізгі – нөлдік (H_0) болжам белгілі критерий арқылы сыналады. Критерийлер – статистикалық және реттік деп екіге ажыратылады. Алғашқысы (статистикалық) кездейсоқ лингвистикалық шаманың

[1] *Фрумкина Р.М.* Вероятность элементов текста и речевое поведение. М., 1971. 168 с.

үлестірілуіне (қолдану жиілігіне) бағыттала отырып, нөлдік болжамды сынауға (тексеруге) қолданылады, ал екіншісі аталған шамалардың реттік үлестірілуін қарастырады.

Әдетте, лингвистикалық болжамдар кездейсоқ таңдама бөліктер негізінде тексеріледі. Бірақ бұл тілдік бірліктің бас жиындағы үлестірілу көрінісінің дұрыстығының айғағы бола алмайды, себебі, мұндай жағдай «болжамды сынау» процесін жалған жолға бұрып жіберуі де мүмкін. Сондықтан дұрыс шешім орнына, мысалы, дұрыс нөлдік (H_0) болжам орнына, бұрыс (H_1) болжам (не керісінше) қабылданады. Осындай жайттарға байланысты зерттеушінің қай болжамды (H_0 не H_1) негізгі деп қабылдауының көп мәні бар. Осы айтылғаннан жіберілетін қателердің мәнділік дәрежесі екі түрлі типпен ажыратылады. Олар – 1-ші және 2-ші дәрежелік мәнді қателер деп саналады. Тіл білімінде қатенің мәнділік дәрежелерін таңдау, көбінде, субъективті (жеке бастық) ой қорыту негізінде іске асады. Мысалы, түркі, монғол, тунгус-маньчжур тілдерінің «алтайлық гипотеза» деп аталатын генетикалық туыстығы жайлы болжам белгілі статистикалық критерий арқылы тексеруге болатындай тұжырымдалған[2]. Мұндағы нөлдік болжам бойынша барлық алтай тілдері бір ғана генетикалық туыстас жеріне, яғни негіз-тілге (язык-основа) келіп тіреледі (восходят). Осы тұжырымды теріске шығаратын өзіне нағыз кәміл сенетін алтайшы да аса қауіпті 1-ші дәрежелік мәндегі қате жіберген болар еді. Бұл жерде нөлдікке қарсы қабылданған болжамның өзі дұрыс болмаса да, мәнділік жағынан 2-ші дәрежелі қате қауіпсіздеу деуге болады. Сондықтан, болжамды тексеру (сынау) критериясын таңдау, сонымен бірге сынау аясын таңдаумен ұштасады деуге болады.

Сөйтіп қандай да болсын дұрыс шешім қабылдар тұсында 2-ші типті қате жіберуден гөрі 1-ші типті қате жібермеу аса маңызды деп саналады. Қате дәрежесінің мәнділігін азайту мақсатымен болжамдар өзара тексерілуі (сыналуы) қажет.

Тәжірибелік құндылығына қарай, яғни 1-ші типті қатеден күтілетін салдарына қарай, зерттеушіні қанағаттандыратын маңыздылық деңгейі таңдалады. Басқаша айтсақ, жіберген қатеден туындайтын салдардың сипатына қарай маңыздылық деңгейін де төмендеткен жөн болады. Әдетте, лингвистикалық зерттеулерде маңыздылық деңгейінің $\alpha=0,05$ немесе $\alpha=0,01$ шамасы да жеткілікті деп саналады. Мәселен, алтай тілдеріне қатысты болжамды статистикалық тәсілмен тексеру (сынау) үшін көрсетілген маңыздылық деңгей шамасы қанғаттандырарлық деп саналады.

Лингвистикалық статистика қазір өзінің алғашқы кезеңін, яғни лингвистикалық құбылыстарды сипаттауды аяқтап, келесі күрделі кезеңіне өтуде. Бұл – тілдік заңдылықтарды сандық тұрғыда алдын ала болжай алатын теория құру мәселесі.

Сонымен, лингвистикалық бірліктердің мәтін ішінде үлестірілуінің

---

[2] *Клоусон ДЖ.* Лексико-статистическая оценка алтайской теории//ВЯ, 1969. №5, С.22-41.

математикалық тұрпатын анықтау мәселесі тілдік құбылыстың моделін құрудағы ең бір маңызды іске айналып отыр[3].

Аталған мәселе бойынша қазақ тілі мәтіні ішінде сөз таптарының статистикалық үлестірілуінің кейбір теориялық заңдылықтарға сәйкес келу болжамын сынауға қатысты зерттеулер жүргізілді. Теориялық үлестірілу заңдылықтары ретінде математикалық статистикадағы «нормальды үлестірілуі», «Пуассон үлестірілуі» және «Шарлье үлестірілуі» заңдылықтары қарастырылды.

Қазақ тілінің мәтіндік нысаны ретінде М.Әуезовтің «Абай жолы» романынан 10 мың *сөзқолданыс*, газет мәтінінен 50 мың *сөзқолданыс* таңдама бөліктер тәжірибе ретінде алынды. Ал сөз таптары ішінен ең негізгілері деп: *зат есім, етістік, сын есім* және *үстеу* сөздердің мәтін ішіндегі статистикалық үлестірілуі сынаққа түсті.

Белгілі бір лингвистикалық бірліктің (мысалы, белгілі бір сөз табының) тілдегі қолданылу сипатын танып білу кезінде құбылу заңдылығы бұзылмайтын ең кіші (минималды) мәтін көлемін анықтау зерттеу ісінің маңызды шарттарының бірі.

Осындай мақсатпен жүргізілген тәжірибе реті төмендегідей:

1) мәтін ішіндегі сөздер сөз табына ажырату және оларды шартты белгілермен сәйкестендіру;

2) мәтіннің жалпы көлемі 25, 50, 100, 200 және 500 сөзқолданыстан тұратын микробөліктерге, яғни сериялық таңдама бөліктерге бөлу;

3) өзгермелі шамаларға қатысты шартты белгілеуді ұстану: әр стиль бойынша алынған мәтіннің жалпы көлемі – N; микробөлік (серия) көлемі – k; әр ерияға сай микробөлік саны – n.

4) әр жағдайдың жалпылама жазылу пішінін қабылдау – $B(N/k)=n$ (теңдік белгісі шартты). Мұндағы $N=n\cdot k$.

Мысалы, «Абай жолы» романы мәтіні бойынша алынған 10000 сөзқолданыстан тұратын мәтін көлемін үшке бөліп: $N_1=2500$, $N_2=5000$, $N_3=10000$ қарастырсақ, әрбір N_i көлемді $k=25, 50, 100, 200, 500$ сөзқолданыстан тұратын микробөліктерге (n) бөлсек, әр бөлікке сай, олардың саны да өзгеріп отырады. Сол сияқты, газет мәтіні бойынша тәжірибеге алынған *50000 сөзқолданыстан* тұратын мәтін көлемі де осылайша серия бөліктеріне бөлініп зерттелді.

5) әрбір вариантқа қатысты микробөліктер ішіндегі қажетті сөз табының жиілігі саналып, бірдей жиілікпен кездесетін сөз табының серия саны анықталды. Осы деректердің негізінде жиілік пен микробөлік санынан тұратын дискретті вариациялық үлестірілімдік қатар түзілді.

Жүргізілген тәжірибе бойынша, көркем әдебиет пен газет мәтіндеріндегі негізгі сөз таптарының мәтін бойында таралуының вариациялық қатарларын (статистикалық үлестірілуінің) кейбір теориялық заңдылықтармен үйлесімділік деңгейі зерттелді. Теориялық заңдылықтар

---

[3] *Фрумкина Р.М.* О законах распределения слов и классов слов // Структурно-типологические исследования. М., 1962. С. 124-133.

ретінде Пуассонның, Шарльенің *(А, В түрлері)* және нормальдық, логарифмды-нормальдық үлестірілу заңдылықтары алынды.

Аталған теориялық үлестірілімдік заңдылықтардың *Пуссон* және *Шарлье-В түрі* дискретті вариациялық қатар құру арқылы, ал *нормальды, логарифмды-нормальды* және *Шарлье-А түрі* үздіксіз вариациялық қатар құру негізінде олардың теориялық жиіліктері мен статистикалық параметрлері анықталды.

Сонымен, құрылған дискретті және үздіксіз вариациялық қатарлар мәтінге қатысты сөз таптары жиіліктерінің эмпирикалық үлестірілімдік заңдылығын анықтайды. Арнайы компьютерлік бағдарламалар негізінде теориялық жиілік пен статистикалық параметрлер есептеліпп шығарылуы қажет.

Енді, осындай зерттеуді, яғни эмпирикалық үлестірілудің таңдалып алынған теориялық заңдылықпен сәйкестігінің мәнді не мәнсіз екендігін анықтау үшін жүргізілетін тәжірибе реті төмендегідей[4]:

1. Зерттеуге тиісті мәтінді және тілдік бірлікті таңдау. Мысалы, қазақ көркем әдебиет мәтіндеріндегі зат есім сөз табына қатысты сөздер.

2. *B(N/k)=n* көрінісіндегі таңдалып алынған параметрлеріне *(N, n, k)* сай етіп, зерттелетін мәтінді таңдама бөліктерге бөлу.

3. *N, n, k* параметріне және тілдік бірлікке сәйкес жиілік сөздіктер түзу.

4. Жиілік сөздіктер негізінде дискретті не үздіксіз вариациялық қатар құру.

5. Тиісті математикалық өрнекті (формуланы) пайдаланып, жоғарыда сөз болған мына параметрлерді есептеп шығару:

а) математикалық күтілуді *(а)*;

ә) ауытқудың квадраттанған ортаны *(* σ *)*;

б) ассиметрия *(А)* мен эксцесті *(Е).*

в) тілдік бірліктің теориялық жиілігін *(* $n_i^T = N \cdot P_i$ *)* ;

г) әрбір *i* тобына қатысты «хи квадрат» *(* χ^2 *)* мәнін.

6. Жиіліктің *5-тен* кіші мәндерін қажеттілікке сай біріктіру және $\sum_{i=1}^{k} \chi_i^2$ қосындысына қажетті өзгертулер енгізу (қолмен).

7. Еркіндік дәреже санын анықтау. Мысалы, үлестірілімнің Пуассон заңы үшін $\ell=1$, нормаль мен логнормаль, Шарлье В заңдары үшін $\ell=2$, Шарлье А үшін $\ell=4$.

8. Есептеліп шығарылған χ^2 пен еркіндік дәреже саны *(*ν*)* бойынша $p(\chi^2 \geq \chi_q^2)$ – ықтималдық интегралдың кестелік мәнін табу[5].

Лингвистикалық зерттеулерде сериялар саны мәтіндегі сөзқолданыстар санынан басқа, кітаптың жол саны не бет санымен де өлшенуі мүмкін. Сол сияқты, параметрлерді нормалау кезінде мәтін бірліктерінің қайсысы және

---

[4] *Джубанов А.Х.* Квантитативная структура казахского текста (опыт лингвистического анализа на ЭВМ). Алма-Ата, 1987. С. 147.

[5] *Романовский В.И.* Применение математической статистики в опытном деле. М.–Л., 1947. 247с.

қандай көлемде алынатыны алдын ала нақтылануы қажет. Әрбір нормалау сипатына сай және өзіне ғана тән теориялық үлестірілу заңдылығы сәйкес келуі мүмкін.

Мәселен, нақты бір нормалауда сынаққа алынған тілдік бірлік «Пуассон үлестірілу» заңына, ал басқаша түрдегі нормалауда сол бірлік мәтін ішінде «Нормаль үлестірілу» заңына «бағынуы» немесе белгілі бір ішкі бөлік көлемінен бастап тек бір ғана теориялық заңдылыққа сәйкес келуі мүмкін екені анықталды[6]. Сондықтан үлестірілу заңдылықтарының параметрлерін салыстыруда және оған лингвистикалық тұрғыдан түсініктеме беруде зерттеу нысанын (мәтінді) нормалау сипатына аса көңіл бөлінуі қажет.

Сонымен, мәтінді ішкі бөліктеудің ең тиімді жолын табу лингвистикалық модельдеу ісінде маңызды деп саналады.

Енді қазақ мәтіндеріне қатысты жүргізілген кейбір тәжірибе нәтижелеріне тоқталайық[7].

Сөз таптарының тілде қолдану жиілігін сөз еткенде, барлық тілдерге де ортақ заңдылық – зат есім мен етістік сөздердің өнімді қолданылуы. Мысалы, М. Әуезовтің «Абай жолы» романы мәтінінде зат есім 35,36%, ал етістік сөздер 31,44%, яғни 100 пайыздың 70-не жуығы тек осы екі сөз табына қатысты сөздер екен. Газет мәтінінің 45% – зат есім, 22 пайызы – етістік (Ахабаев).

Сондықтан қазақ мәтіндеріндегі сөз үлестірілу заңдылығын зерттеу жұмыстарына алынатын таңдама мәтіндер (выборки) әр сөз табына байланысты осындай қарапайым және тәуелсіз талаптарға жауап берулері керек. Егер аталған шарттар орындалды деп ұйғаруға болса, зат есім, етістік сөздердің (не басқа сөз таптарының) тәжірибелік (эмпирикалық) үлестірілудің теориялық (нормальды не басқа) заңдылыққа үйлесімді келеді (согласуется) деп *«H_0»* жорамалын жасауға болады. Бұл жорамалды «бағалау» әр түрлі *N, n, k* шамаларына қатысты $p(\chi^2 \geq \chi_q^2)$ ықтималдықтарды салыстыру негізінде іске асады.

Көркем әдебиет мәтіні, яғни «Абай жолы» романынан тәжірибе ретінде алынған $N=50000$ сөзқолданыс:

а) $N_1=2500$, $N_2=5000$, $N_3=10000$, $N_4=20000$, $N_5=25000$, $N_6=40000$ және $N_7=50000$ көлемдері ретінде бөлек-бөлек қарастырылды;

ә) ішкі бөлік көлемі *(n)*: $n_1=100$, $n_2=200$, $n_3=250$, $n_4=500$, $n_5=1000$ сөзқолданыстан тұратын көлемдері ретінде бөлек-бөлек қарастырылды;

б) барлық жағдайда $N=k\times n$ болатынын ескеріп, сериялар шамасы $k=N:n$ анықталды: $k_1=25$, $k_2=50$, $k_3=100$, $k_4=200$, $k_5=500$ сериялық сан мөлшері екені анықталды.

---

[6] *Бектаев К.Б., Зубов А.В., Ковалевич Е.Ф.* и др. К исследованию законов распределения ленгвестических едениц // Статистика текста. Минск, 1969. С. 131-162; *Лукьяненков К.Ф.* Использование схем Пуассона и Гаусса в исследовании распределения лингвестических единиц текста // Вопросы лингвостатистики и автоматизация лингвистических работ. М., 1970. Вып. 2. С. 3-14; *Феллер В.* Введение в гориювераятностей и ее приложения. М., 1967. Т.1. 498 с.

[7] *Джубанов А.Х.* Квантитативная структура казахского текста (опыт лингвистического анализа на ЭВМ). Алма-Ата, 1987. С. 147.

Осылайша нормалаудың нәтижесінде көркем әдебиет мәтіні бойынша зат есім сөздердің тәжірибелік жиіліктері оның теориялық жиіліктерімен сәйкес келудің «сенімділік дәрежесін» қанағаттандыру жағын алғанда 1-ші орында «Нормаль үлестірілу» заңы, ал 2-ші орында «Шарлье А үлестірілу» заң түрі екені анықталды. Газет мәтіні бойынша, сериялар саны 100-ден артқан жағдайда зат есім сөздерің теориялық заңдылықтарға өте сәйкес келу мүмкіндігі «логарифмды-нормаль үлестірілу» заңдылығы екені дәлелденді.

Көркем әдебиет және публицистика мәтіндеріндегі етістік сөздердің тәжірибелік үлестірілу деректерінің теориялық үлестірілу заңдарымен үйлесімділігін зерттей келе, олардың зат есім сөздерге сәйкес математикалық моделі (үлгісі) нормаль заңдылығы екені айқындалды. Ал егер, серия саны $k \leq 100$ болған жағдайда, етістік сөздердің үлестірілулері теориялық заңдылықтар ішінен «Шарлье А үлестірілу» типіне сәйкестігі айырықша байқалды.

Публицистикалық мәтіндердегі етістікке қатысты сөздердің тәжірибелік үлестірілуі параметрлерді нормалаудың мынандай түрінде: $\{N \geq 1000,\ n \geq 100,\ k \geq 100\}$, теориялық заңдылықтар ішінен «нормаль үлестірілуін» көбірек қанағаттандыратынын ықтималдықтың дәрежелерінен байқауға болады.

Сонымен, қазақ тіліндегі етістікке қатысты сөздердің тәжірибелік жиіліктерінің мәтін бойындағы үлестірілуі параметрлерді аталған нормалауға сай, «Нормаль заңымен» үйлесімдік табады деп қорытындылауға болады.

Теориялық үлестірілу заңдарының Пуассон және Шарлье В типі қазақ тіліндегі *зат есім* мен *етістік* сөздердің тәжірибелік жиілік үлестірілу деректерімен үйлеспейтіндігі де байқалды.

Зат есім мен *етістік* сөздердің қатарында сын есім сөздердің де мәтіндегі үлестірілуі қарастырылды. Қысқаша айтқанда, олардың теориялық заңдылықтарымен сәйкестігі көркем әдебиет пен газет мәтіндерінде бірдей емес және нормалау түріне қарай, яғни N, n, k параметрлерін бөлшектеуге қатысты айырым табады. Дәлірек айтқанда, көркем әдебиет мәтіні үшін нормалау сипаты – $\{N \geq 2500,\ n \geq 1000,\ k \geq 25\}$ болғанда, «Пуассон үлестірілу» мен «Шарлье үлестірілу» заңдылығының А мен В типтерімен бірдей дәрежеде үйлесетіні байқалды, ал газет тіліне қатысты $\{N \geq 5000,\ n \geq 100,\ k \geq 50\}$ нормалау әдісі ғана аталған заңдылықтарды қанағаттандырады екен.

Үйлесімдік дәрежесін оның «артықшылық» тұрғысынан қарастырсақ, әрине, *сын есім* сөздерге қатысты Пуассонның теориялық үлестірілу заңы бірінші орынға ие. Солай бола тұра, нормалаудың $\{N \geq 2000,\ n \geq 100,\ k \geq 200\}$ мәндерінен бастап, қазақ тілінің *сын есім* сөздері «нормаль заңдылығымен» де жақсы үйлесімдік табатынын байқаймыз.

Сөз таптарының тәжірибелік жиіліктерінің теориялық үлестірілу заңдылықтарымен үйлесімдігі, көбіне, сериялар санына қатысты екені байқалады. *Үстеу* сөздерге қатысты ең «қолайлы» деп Пуассон үлестірілу заңын атауға болады. Ал сериялар саны $k > 500$ болғанда үлестірілу «нормаль заңымен» үйлеседі.

Сонымен, жоғарыда баяндаған жайттар мәтінді мөлшерлеу (нормалау) әдістерінің қазақ тіліндегі сөз таптарының тәжірибелік (эмпирикалық) үлестірілулерінің белгілі бір теориялық үлестірілу заңдылығына сәйкес келу не келмеу мүмкіндіктерін айқындауды көздейді. Бұл зерттеудің нәтижелері қазақ тіліндегі құбылыстарды модельдеу мәселелерінде аса маңызды рөл атқаратынына сеніміміз мол.

ҚОЛДАНБАЛЫ ЛИНГВИСТИКА ПӘНІ ЖӘНЕ ОНЫҢ НЕГІЗГІ БАҒЫТ-БАҒДАРЫ

Қазақстан Республикасы Білім және ғылым
минстрлігі А.Байтұрсынұлы атындағы Тіл білімі
институты «Академик Рабиға Сыздық және қазақ
тіл білімінің мәселелері» атты халықаралық ғылыми-
теориялық конференция материалдары. «Арыс»
баспасы. Алматы-2004. 179-183-бб.

Тілдің қолданбалы аясы бұрыннан да өзінің кең және жан-жақтылығымен ерекшеленетін. Оның ескіден келе жатқан саласы – жазу (графика), ана тілі мен шет тілін оқыту әдістемесі және тілдің лексикографиялық жүйесі. Осылардан кейін барып – аударма, шифртану (дешифровка), орфография, транслитерация және терминологияны өңдеу. Қолданбалы лингвистиканың дәстүрлі бағыттарының бірі – мемлекеттің тіл саясатына қатысу: 1) әліпбиді, орфографияны өңдеу, сауатсыздықты жою; 2) мемлекеттік тілді тағайындау; 3) мемлекеттік тілден басқа тілдердің орнын анықтау; 4) ұлттық терминологияны өңдеу, бір ізге келтіру және қалыптастыру; 5) қала, көше, алаңдарға ат қою мен оның бұрыннан келе жатқан атын өзгерту мәселелері, яғни ономастика мәселелері.

Қолданбалы тіл білімнің осындай классикалық бағыт-бағдарын дамыту және мейлінше жетілдіру мәселелерімен бірге XX ғасырдың екінші жартысынан бастап, оның бір қатар жаңа бағыттары орын ала бастады. Бұл бағыттар заманға сай қоғамдық, жаратылыстану және техникалық ғылымдардың өзара даму сипатынан туындайтын қолданбалы лингвистиканың тарихи дамуының логикалық жалғасының көрінісі еді. Негізінен алғанда, адам баласының қызметінің әртүрлі аясындағы жұмыстарды жеңілдету бір ғана проблемалық мәселеге тіреледі. Ол – қоғам өміріндегі ақпарат жұмысын өңдеу. Бұндай атқарат мәтін түріндегі жазба тілінде не үйреншікті сөйлеу тілінде болуы мүмкін.

Қоғамның практикалық қажеттігін өтеуде, әсіресе, ақпаратты сақтау мен оны тарату істерінің сырын ашуда зерттеуші оны ең алдымен тілдің ішкі табиғатындағы зандылықтардан іздегені жөн.

Осымен байланысты өзекті мәселенің бірі – оператордың дауысымен басқарылатын станок пен құрал-аспаптарды өндіру. Ол үшін әртүрлі тілдердегі ақпараттарды автоматты түрде өңдеу мен оларды іздестіру,

байланыс торабы жұмысын (телефон, радиобайланыс т.б.) жетілдіру, адам баласының сөйлеу мен есту қабілетінің бұзылуына қатысты ауруларды емдеуде дыбыстың фонетикалық деректеріне жүгіну және т.б.

Аталған проблемалық мәселелерді ЭЕМ-ды (компьютерді) кең түрде қолдана отырып, мәтінді (жазба, сөйлеу) автоматты түрде өңдеу. Мәселен, әр типті ақпараттық жүйелерді тілдік деректермен қамтамасыз ету; машиналық аударма мәселесі; табиғи тілді түсінетін жүйе құру (жасанды интеллект жүйесіндегі тілдік мәселелер); сөйлеу кезіндегі дыбыстық сигналдарда бар ақпараттарды пайдалану негізінде арнайы жүйе құру т.б.

Соңғы жылдары әр түрлі ақпараттық жүйелерді тілдік деректермен қамтамасыз етумен байланысты терминжасау мәселесін біріздiлікке түсіріп, тұрақтындыру өзекті мәселеге айналып отыр. Себебі, әр білім саласының ғылыми және техникалық терминдерді қолдану қажеттігі барынша сұраныс тудыруда. Әрине, мұндай мәселе лексикография саласынан тыс өз шешімін таппайтыны белгілі. Сондықтан қоғам өмірінде ЭЕМ-нің (компьютердің) кең қолданылуымен тығыз байланысты автоматты лексикография саласы біртіндеп өз отауын құруда.

Практикалық мәселелерді шешуге байланысты табиғи тілде диалогті іске асыру үшін адам мен ЭЕМ арасында әсерлі қарым-қатынас орнату қажеттігі туып отыр. Мәселен, оларға жататындар: сұрақ-жауап жүйесін құру, роботтарды басқару жүйесін құру, аталған басқару жүйелерінде дұрыс шешім қабылдау үшін диалогтік процесті іске косу. Мұндағы ең негізгі мәселе – автоматты құрылғының жазба не дыбыстық мәтінді түсіне білу проблемасы.

Әрбір тілдің қолданбалы мәселені шешуде өзіне тән ерекшелігі бар. Мұндай есептердің санын, түрін алдын-ала білу мүмкін емес. Өмірде олар жыл сайын, ай сайын ауысып, біреуі келіп, екіншісі кетіп жатады. Қолданбалы лингвистикаға қатысты іргелі зерттеулер ғылыми-техникалық, ұйымдастыру-басқарудың ауызша-жазбаша күй жаттарының, сөздік түзу мен және т.б. түрлерінің фонетикалық, грамматикалық, семантикалық және статистикалық құрылымын сипаттау мен модельдеу жақтарын қамтиды. Жекелеп айтсақ, оған теориялық және қолданбалы лингвистикалық шекарасында жатқан мәтіндік бірліктердің формалды моделін (үлгісін) құру проблемасын жатқызуға болады.

Теориялық линвистика, негізінен, тілді қалыптық, жүйелік тұрғыда қарастыратындықтан, қолданбалы тіл білімі тілді әрекет үстінде, яғни оның қарым-қатынас (коммуникация) кезіндегі табиғатын түсінуге тырысады. Көптен бері қолданбалы тіл білімі ғылыми-техникалық пен іскерлік прозадан өзін алшақ ұстағаны аян. Дегенмен, XX ғ. 70-жылдарында ғалымдардың мынадай шешім жасауына тура келді. Көптеген қолданбалы мәселелер таза лингвистикалық емес, тіпті дәлірек айтсақ, ондай проблеманы шешу адамның іс әрекеті мен ойлау процесін, тілдің семантакасы мен оның формалды және семантикалық әдіс-амалын синтездеу (жинақтау) жолдарын модельдеуге тікелей қатысты. Осыдан барып қолданбалы тұрғыдағы

зерттеулердің іргелі проблемасы – білімді модельдеу (моделироваиие знаний) мәселесі анықталды.

Соңғы кезде бұл проблема (білімді модельдеу) бірнеше ғылымдар тоғысына, дәлірек айтқанда – логика, лингвистика, математика, психология, кибернетика салаларының аясына қатысты екені анықталды. Көптеген ғылымдардың осы мәселеге қызығушылығының негізгі себебі біреу-ақ, ол – автоматты жасанды интеллект жүйесін құру. Солай бола тұра, «білім» бізге «тіл» арқылы берілетіні анық. «Білім» сөйлеу мәтінінде де (монолог, диалог, әр жағдайдағы реплика) және сонымен бірге жазба мәтіндерде де (ескі жазба ескерткіші, көркем әдебиет, ғылыми-техникалық әдебиет) көрініс табады. Тіл арқылы «білімді» біз ұрпақтан ұрпаққа жеткіземіз. Сондықтан, «тіл» «білімді» сақтау формасы және оны тарату құралы десек те болады. Біздіңше, ғылым мен техникада мәтіннен тыс жатқан «білім» жоқ деуге болады, ал ғылыми-техникалық мәтіннің семантикасын модельдеу дегеніміз – ол осы салалардың білім жүйесін модельдеу. Осылайша, біртіндеп келіп, білімді модельдеу проблемасы мәтін мазмұнын (мағынасын) модельдеумен ұштасып жатқанына көз жеткізуге болады. Бұл жерде ең негізгі шешімін табуды қажет ететін мәселе – мәтіннің семантикалық көрінісін (семантическое представление) құру.

Дегенмен, техникалық және іскерлік коммуникация аясындағы нақты қолданбалы мәселені шешу сол саладағы құжаттық мәтіндерінің грамматикалық, лексикалық, семантикалық құрылымының сипатталуына байланысты. Сонымен бірге, аталған салалардың терминологиялық сөздігінің болуына, мәтіннің статистикалық құрылымының зерттелуіне және осындай мәтіндер типтерінің семантикалық көрінісінің толық құрылуына да байланысты екенін айта кету қажет.

Қолданбалы лингвистиканың дамуы, яғни ғылыми жетістіктерге ие болуы тіл білімі теориясының дамуына, да өз әсерін тигізбей қоймады. Мысалы, XX ғ. 20-30-жылдары практикалық қажеттіктен туындаған ғылыми-техникалық терминологияны бірізді және тұрақты ету жолдары жаңа лингвистикалық пән – «терминтану» (терминоведение) саласын өмірге әкеледі. Сол сияқты, Кеңес одағы халықтары тілінің әліпбиі мен жазуын құруға байланысты 30-40-жылдары жүргізілген кең көлемді әрі орасан күрделі, әрі зор тәжірибелік мәні бар жұмыстар тілдерді синхронды сипаттаудың әдістерін жетілдіруге себепші болды (ынталандырды) деуге болады.

Соңғы 20-30-жылдарда пайда болған қолданбалы лингвистиканың жаңа аспектілерінің бәрі бір ғана ортақ проблемаға тәуелді екенін байқатты, ол – тілдің жазба не сөйлеу түрлерін автоматты өңдеу мәселесі еді. Мұндай проблемамен айналысу тілді талдау мен сипаттаудың жаңа әдіс-тәсілдерінің дамуына және тіл табиғаты мен тіл білімі құрылымында деген жаңаша көзқарастың пайда болуына мүмкіндік туғызды.

XX ғасырдың 50-70-жылдарында тілдік материалдарды автоматты түрде өңдеуге қатысты күрделі проблеманың шешімі табылуы қолданбалы лингвистиканың әрі қарай дамуына айтарлықтай әсер етті.

Қолданбалы лингвистиканың жаңа аспектілері теориялық тіл біліміне жаңадан ғана ене бастаған математикалық әдістерді, әсіресе, теоретика-жиындық, формалды-логикалық, статистика-ықтималдық әдістердің қолдануын барынша жеделдетті.

Классикалық тіл білімінің классикалық логиканың, психологияның, семантика мен математиканың қатысуынан барып тілдегі модельдеу әдісінің өзінше бөлек ғылым ретінде көрініс табуының нәтижесінде қазіргі «құрылымдық лингвистика» деп аталып жүрген ғылыми сала өмірге келгені белгілі. Сол сияқты, енді тіл біліміндегі дербес тұрған теориялық бағыт ретінде «тілдің теориялық модельдері» атты жаңа тарау пайда болды.

Лингвистика мен математика салаларының аралығынан туындаған – «математикалық лингвистика» атты жаңа пән пайда болды. Бұл пәннің XX ғ. 50-60-жылдары өмірге келіп, қалыптасуы бұрыннан да тіл білімінің барлық ішкі дамуының өзінен де сезіле бастаған болатын. Мәселен, бұл жайт тілдің құрылымына көңіл аударудан, тілдік жүйе ретіндс қарастырудан және оның микрожүйелерден тұру күрделілігін жете түсінуден, лингвистикалық нысандар мен олардың атрибуттары аралығындағы қатынастарды танып білуден байқалып, олар математикалық лингвистика пәнінің өмірге келуіне объективті жағдай туғызды деуге болады.

Атап айтқанда, қазіргі теориялық тіл білімінің дамуына ең көп әсер еткен – құрылымдық лингистика мен математикалық лингвистика пәндері. Бұндай әсер көбінде грамматиканың синтаксис және семантика салаларында айтарлықтай көрініс тапты.

Құрылымдылық синтаксисте мынадай формалды екі синтаксистік модель жүйелі түрде жете зерттелді: тікелей құрастырушылардың моделі және өзара байланыстылықтың моделі. Екі модель де машиналық аударма жасауда, автоматтанған синтаксистік талдау мен мәтінге жасалатын басқа да автоматты өңдеулерде кең түрде пайдаланылады.

Семантика тілдің барлық деңгейлеріне ортақ болғандықтан тіл білімінде оны бөлек деңгей ретінде қарастырмайды. Семантикада, біріншіден, лингвистиканың өзіне ғана тән әдістің әсері байқалса, екіншіден, логикалық семантика мен теориялық классификациялау әдістерінің де әсері орын алады. Синтаксистік семантикада сан жағынан көбірек зерттелген жайттар – сөйлемнің семантикалық құрылымы жайлы тұжырымдар (концепциялар).

Қоғамдық ғылымдарда ғылыми зерттеулерді автоматтандыру, ең алдымен, құжаттық және фактографикалық деректердің ауқымды қорын құру барысында қажет болды. Себебі, мұндай автоматтандыру типі ең алдымен қоғамдық ғылымдар мамандарының өз қалауынан туындап отыр. Мәселен, ғылым жолындағы қандай анықтамалар түрі қоғамдық саладағы қызметкерге қажеттірек. Біріншіден, мұндағы сұраныс кітапханалардағы библиографиялық істе: белгілі бір мәселеге қатысты барлық отандық не шетелдік әдебиеттерді іздестіріп, тауып алу сияқты мәселелер. Мұндай материалдарға жататындар: кітаптар. мақалалар, конференция тезистері, хроникалық шағын мақалалар және т.б.

Екінші кезеңде, жиналған әдебиеттер ішінен қажетті деректерді жинастыру, оларды қажетті белгілеріне қарай сұрыптау, топтастыру және т.б. жұмыстар орындалуы керек.

Көптеген қоғамдық ғылымдарға қатысты мәселелер, әсіресе, тіл білімінде, эгнографияда, антропологияда, тарихта карталық мәліметтерге тікелей қатысты болып келеді. Сондықтан ЭЕМ-ге ондай деректерді енгізу мен қағаз бетіне шығару өте-мөте өзекті деп саналады.

Қоғамдық ғылымдар бойынша ғылыми-зерттеу жұмысын автоматтандыруды қажет ететін мәселелер:

а) әр тілдер бойынша белгілі такырыпқа қатысты әдебиеттерді іздеп, табу;

ә) жазба ескерткіштер мен мәтіндік материалдар қоры бойынша таңдама жұмыстарын атқару;

б) табылған материалдар арқылы қажетті амал-әрекеттерді іске асыру;

в) материалды жан-жақты сұрыптау;

г) алдын ала белгілі өлшемдер бойынша реестрлер, каталогтар сипаттамаларын жинақтау;

д) статистикалық, картографикалық, сұрыптау теориясы мен жүйелік талдау әдістерін қолдану;

е) берілістерді (деректерді) сызба, сурет, карта түрінде көрсету.

Сонымен, ғылыми жұмысты автоматтандыру процесінде құрастрылған жүйе негізінде тұтынушы маманға қажетті тарихи деңгейде зерттеу нысанына қатысты барынша толық энциклопедиялық білім және библиографиялық мәліметтер берілуі қажет деп санаймыз,

Жоғарыда сөз болған мәселелерді қорыта келе айтатын болсақ, қолдалбады лингвистика, ең алдымен, кешенді пән. Себебі, ол философия, психология, физиология, математика, логика, әлеуметтану, информатика салаларымен үнемі тығыз қатынаста болуы керек.

Жаңаша мағынадағы қолданбалы лингвистиканы жеке қарастыратын болсақ, ол көптармақты ғылыми сала болып табылады. Ғылыми әдебиеттерде оны «компьютерлік лингвистика», «инженерлік лингвистика», «автоматты лингвистика», «есептеу лингвистикасы», «квантитативті лингвистика», «статистикалық лингвистика» деген жарыспалы терминдермен атап жүр. Әриие, бұл атауларға сәйкес салалардың өздеріне тән бағыт-бағдары ажыратылады, алайда бәріне тән ортақ мақсат – адам баласының қарым-қатынасы үшін қажетті табиғи тіл қызметінің ең ыңғайлы, ең тиімді жолдарын іздестіру.

Қолданбалы лингвистака салаларының барлық тілдерде, соның ішінде қазақ тілінің жазба түріне қатысты зерттеулерде ең көп қолданыс тапқан тармағы – статистикалық лингвистика. Дәстүрлі әдіспен тілді зерттейтін тілші-ғылымдардың бәрі бірдей статистикалық әдіс-тәсілдерді қолдай бермейді, бірақ өз ізденістерінде осы саланың ***көп, аз, мол, жиі, сирек, өнімді, өнімсіз*** тәрізді терминдерін барынша пайдаланады және олардың статистикалық ғылым саласына қатынастылығын көп жағдайда аңғара бермейді.

Статистикалық лингвистика тілді зерттеудің тек әдісі не тәсілі болып қана қалмай, қазіргі кездегі тілтануда өзінше бір бөлек ғылыми пән дәрежесіне көтеріліп отыр. Бұл пән тілдік бірліктердің сапалық және сандық (мөлшерлік) мәліметтерін тілдің табиғатына сай қарастыруды жөн санайды.

Статистакалық лингвистика, негізінен, математикалық статистика әдістемесіне сүйенеді. Зерттеу барысында ондай әдістемені қолдану қатынас құралы ретіндегі «тілді» белгілі бір жүйе деп қарастырудан туындайды. Шынында да, «тіл» дегеніміз объективті болмыста өмір сүретін көп өлшемді және белгілі бір ішкі заңдылықтардың негізінде топталған (реттелген) табиғи құбылыс деуге болады. Ал тілдік жүйе математикалық статистика заңдылықтарына бағыну үшін ондағы бірліктер тобы қайталанып отыратын және кездейсоқтық сипатта болатын тілдік элементтерден тұруы қажет. Бұл жерде айта кететін жайт, тіл қызметі кезінде (яғни тілдік қарым-қатынас кезінде) дыбыстар, сөздер, сөз тіркестері, сөйлемдер және одан да үлкен бөліктер қайталанбайтын сипатта болса, адамдар арасындағы тілдік қатынас бұзылып, ақпарат таратушы мен қабылдаушы арасында түсінбестік пайда болған болар еді. Сондықтан да тілдік бірліктердің қайталануы олардың тілімізге тән болуынан, олардың қолдану сипатына ***жиі, сирек*** деген ұғымдардың да тән екендігінен туындайды.

Табиғи ортадағы элементтер арасындағы қайталану сипатының негізінде математикалық статистика мен ықтималдықтар теориясы атты ғылым салаларында тілдік бірліктердің статистикалық заңдылықтары қалыптасады. Зерттеуге алынған бірліктің (единица, элемент) ондай заңдылыққа бағыну не бағынбауын айқындау үшін алынатын ортаның аумағы (мәселен мәтін көлемі) барынша мол болғаны жөн. Мысалы, кітаптың бір бетінде кездесетін сөздер мен оның жүз бетінде кездесетін сөздердің, қайталану сипаты әр түрлі болатындығына ешкімнің күмәні жоқ. Ал заңдылықты айқындау үшін, кітаптың бір бетіндегі мәліметтен гөрі оның көптеген беттерінен алынған мәліметтердің шамасы шындыққа жақын болатындығы анық.

Тілге тән заңдылықтарды статистика тәсілімен анықтау негіздемесі мынада:

1) сандық (мөлшерлік) құбылыс тіл табиғатына әуелден-ақ тән болуы;

2) тіл құрылымындағы сандық және сапалық сипаттардың өзара байланыста болуы;

3) тілдің әр түрлі бірліктері сөйлеу ағымында статистикалық зандылықтардың ең болмағанда біреуіне бағынуы тиіс деп ұйғару керектігі.

Әдістердің иңдуктивті, дедуктивті болып ажыратылатынын ескере отырып, негізінен алғанда, статистикалық әдіс – индуктивті әдіс арқылы, ал тілдің әр түрлі модельдерін жасау үшін қолданылатын логика-математикалық әдістер – дедуктивті әдіс арқылы жүзеге асатыны анық.

Тілдің ықтималды-статастикалық моделін жасауға болатыны белгілі десек, ондай модель, бірішіден, белгілі бір тілдік жүйенің мәтіндері арқылы жасалады. Ал екіншіден, тілдің логика-математикалық моделі сол тілдің толық жүйесін айқындай алады.

Сөйлесу арқылы қатынас жасау кезіндегі тілдік процесті *байланыс өзегі* (канал связи) не *байланыс жолы* деп есептеуге болады. Ал мұндай байланыс жолы арқылы таратылатын ақпарат әріп, дыбыс морфема және т.б. лингвистикалық бірліктер негізінде іске асады. Бұл жердегі лингвистикалық бірліктер белгілі бір «кодтың» символдары ретінде есептеледі. Байланыс өзегіндегі хабарлаушы мен қабылдаушы, яғни айтушы мен тыңдаушы өзара бірдей «кодты» пайдаланулары шарт болуы қажет.

Тіл зерттеу тәжірибесінде статистиканың араласуы бұрыннан да бар десек, ондай әдіс тек жекелеген зерттеушілердің ізденістерінде ғана кездессе, енді статистикалық лингвистика жеке ғылыми пәнге, ғылыми бағытқа айналып отыр. Осымен байланысты статистикалық лингвистака саласының қазақ тіліне қатысты өзіне тән алға қойған теориялық және практикалық мақсаты мен міндеті айқындалуда.

Әдебиеттер:

1. *Бектаев К.Б., Пиотровский Р.Г.* Математические методы в языкознании. ч. I. Алма-Ата, 1973; ч. II. Алма-Ата, 1974.

2. *Беляева Л.Н.* Применение ЭВМ в лингвиститческих исследованиях и лингводидактике. Л,. 1986.

3. *Герд А.С.* Предмет и основные направления прикладной лингвистики// Прикладное языкознание. Санкт-Петербург, 1996. С. 5-14.

4. *Звегинцев В.А.* 1) История языкознания XIX–XX веков о очерках и извлечениях. ч. II. М., 1965; 2) Теорическая и прикладная лингвистика. М., 1968.

5. Знание языка и языкознание. М., 1991.

6. *Зубов А.В.* Основы лингвистической информатики: Учеб. Пособие. ч.I. Минск, 1991; ч. II. 1992; ч. III. 1993.

7. *Зубов А.В., Лихтарович А.А.* ЭВМ анализирует текст. Минск, 1989.

8. Использование ЭВМ в лингвистических исследованиях / Под ред. В.И. Перебейнос. Киев, 1990.

9. Компьютеризация общества и человеческий фактор / Под ред. *А.И. Ракитова. М., 1988.*

10. Лингвистический энциклопедический словарь М., 1990.

11. *Марчук ЮЛ.* Математические методы в языкознании. М., 1990.

12. *Налимов В.В.* Вероятностная модель языка М., 1974.

13. Новое в зарубежной лингвистике. Вып. XXIV: Компьютерная лингвистика. М., 1989.

14. *Пиотровский Р.Г.* 1) Текст, машина, человек. Л., 1978; 2) Инженерная лингвистика и теория языка. Л., 1979.

15. Прикладные аспекты лингвистики. М., 1989.

16. *Рогожникова Р.П., Чернышева Л.В., Кузнецова Е.* Основные направления автоматизации лингвистических исследований. Л., 1988.

17. *Рождественский Ю.В.* Техника, культура, язык. М., 1993.

18. Стуктурная и прикладная лингвистика. / Под ред. А.С.Герда. Л., 1978.

МӘТІН АБЗАЦТАРЫН ЗАТТЫҚ-ЛОГИКАЛЫҚ МАЗМҰН ТҰРҒЫСЫНАН ЗЕРТТЕУ

ҚР ҰҒА Хабарлары. Тіл, әдебиет сериясы.
Алматы, 2004. № 4, 21-28-бб.

Алдыңғы ізденістерімізде ұлттық тілімізді зерттеудің әдістерін заманға сай дамыту үшін, оны бейнелейтін мәтін бірліктерінің сыр-сипатын, өзара байланыстық мәнін және белгілі бір тақырыпқа (темаға) қатынасын ашу мәтін лингвистикасының өзекті мәселелерінің бірі екендігін атап өткенбіз. Сонымен қатар, автордың мәтінді абзацтарға бөлшектеуін заңды құбылыс екенін, яғни абзац мәтіннің синтаксистік және мазмұндық құрылымының ықтималдық моделінің элементі болатындығын сөз еттік. Абзацқа тән қасиеттерді саралаумен бірге, олардың мәтін құрылымына қатысты: ***кіріспе, медиалды*** (ортаңғы), ***қорытынды*** деп айырым табуымен қатар, мазмұндық жағынан – ***баяндау, суреттеу, пайымдау*** деп аталатын функционалды-мағыналық топтарға ажыратылатыны дәлелденді [1, 118-124; 2, 7]. Сонымен, абзац мазмұны арқылы ақиқат болмыстағы шекті нысандардың қасиеттері мен олардың өзара қатынастық мүмкіндігі баяндалатын шағын жағдай (микроситуация) берілетіні белгілі. Мәтін абзацтарын функционалды-мағыналық топтарға ажырата зерттеу арқылы тек мазмұнның жалпы типтік жақтарын ғана анықтай аламыз, ал оны нақтылай түсу абзац мазмұнының заттық-логикалық түсінігіне тікелей қатысты [2; 3].

Мәтін құрылымының статикалық және динамикалық құраушыларының болуын қолдауда біз үшін ақиқат болмыстағы деректер (нысандар) – ***баяндау, суреттеу пайымдау*** абзацтарының қайсысына қатысты аталынғанын және олар қандай ***негізгі тірек сөздер*** (НТС), ***қосалқы тірек сөздер*** (ҚТС) немесе олардан ***басқа да сөздер*** (БС) арқылы берілетіндігін зерттеу аса маңызды болып есептеледі. Осы айтылғандарды ескере отырып, әрбір функционалды-мағыналық типке қатысты «абзацтың» заттық-логикалық мазмұнының анықтамалары төмендегідей берілді [1, 125-126; 2, 7-12].

1. Мәтін абзацтарында дәйекті баяндалған ақиқат болмыстағы іс-әрекет пен қалыпты жағдайларды жан-жақты және жүйелі түрде сипаттайтын, жанды не жансыз нысандарды оқиғаның болған орны мен уақытын айқындауына қарай сараланған (жіктелген) жалпы мәтін мазмұнына қатысты ***негізгі тірек сөздер*** (**НТС**) мен ***қосалқы тірек сөздер*** (**ҚТС**) жиынтығын баяндау абзацтарының заттық-логикалық мазмұны деп атаймыз.

2. Мәтін абзацтарында сөз болған шағын оқиғадағы бірінші және екінші дәрежелі жанды, жансыз нысандардың орны мен уақытын суреттеуіне қарай саралап, жалпы мәтін мазмұнына қатысты ақиқат болмыстық деректерді (белгілер, құбылыстар, оқиғалар, көзқарастар, теориялар және т.б.) түгелдей санамалап атайтын ***негізгі тірек сөздер*** (**НТС**) мен ***қосалқы***

тірек сөздер (**ҚТС**) жиынтығы суреттеу абзацтарының заттық-логикалық мазмұны деп аталады.

3. Мәтін абзацтарында баяндалатын шағын оқиғадағы бірінші және екінші дәрежелі (жанды, жансыз) нысандардың орны мен уақытына байланысты ой тізбегінің жалпы мәтін мазмұнына қатысты қатынасын анықтайтын негізгі тірек сөздер мен қосалқы тірек сөздер жиынтығын пайымдау абзацының заттық-логикалық мазмұны деп атаймыз.

Осы анықтамалардың негізінде қазақ тіліндегі әртүрлі стильдерді анықтайтын мәтін абзацтарын функционалдық-мағыналық және құрылымдық (кіріспе, медиалды, қорытынды) типтеріне қарай ажырата отырып, заттық-логикалық мазмұнына қарай топтап, олардың ішкі түрлерін ажырату арқылы статистикалық зерттеу нәтижелерін қысқаша түрде атап өту мақаламыздың мақсаты болмақ.

Аталған топтау жұмыстары, яғни абзацтарды функционалды-мағыналық жағынан ажырата білу аса ұқыптылықты қажет ететіні анық. Қазақ тілінің үш түрлі стилі ***(ғылыми-көпшілік стиль, газет стилі, көркем әдебиет стилі)*** бойынша жүргізілген зерттеу жұмыстары алғашында абзацтардың мәтін бойында орналасу тәртібіне қарай, яғни құрылымдық тұрғыда қарастырылды.

Ғылыми-көпшілікке арналған 47 мәтін негізінде жүргізілген зерттеу жұмысының нәтижелері [1, 126-133] бойынша (Көшербаев К. Минералдар әлемінде. – Алматы, 1986; Қалыбеков Қ. Экология және ашық кен. – Алматы, 1988) бастапқы абзацтар саны 93 кіріспе абзацқа тең деп танылды. Жоғарыда келтірілген анықтамаларға сай, олардың заттық-логикалық мазмұны бойынша талдау нәтижесінде 20 типтік тұжырымдаулар (формулировки) топталып алынып, топтамалардың статистикасы анықталды және мазмұнды формалдау жұмысының қажеттігіне орай, оларды шартты белгілермен кодтау принципі шешілді. Аталған зерттеу нәтижелері арнайы кесте арқылы көрініс тапты. Кестелер 5 бағанадан тұрады: 1 – типтердің рет саны; 2 – бастапқы абзацтар типінің тұжырымдалуы; 3 – типтің коды (шартты белгісі); 4 – типтің абсолютті жиілігі; 5 – типтің қатынастық жиілігі.

Заттық-логикалық мазмұны тұрғысынан топталған мәтін басындағы абзацтардың типтері қолданылу жиілігі жағынан біркелкі болмайтыны айқындалды, сондықтан мақалада олардың жоғары жиіліктегілері (9 және одан жоғары) ғана алынып, ***1-кесте*** үзінді түрінде берілді. Кестеде көрсетілген тұжырымдаулар – қарастырылған ғылыми-көпшілікке арналған мәтіндерге тән мазмұндық типтер. Мәтін мазмұны – тұжырымдауларда аталған басты нысандардың төңірегінде шоғырланып келіп, солардың мән-мағынасын ашуға қатысты автордың (зерттеушінің) өткен, осы шақтарда істелген және болашақта жүзеге асатын іс-әрекеттерінің нәтижелерін баяндау түрінде көрініс табады.

Мысал ретінде осындай ғылыми-көпшілікке арналған «***Кенді ауыр суда байыту***» тақырыбы бойынша жазылған мәтіннің кіріспе абзацтарының бірі алынды: *«Тығыздығы 1 кг/м^3 ден асатын кез келген зат суда батады. Мұндай заттардың өте ұсақ түйіршіктерінің өзі суда қалқуы мүмкін емес.*

Ал мына көріністің мәні ерекше: жұдырықтай кесектер суда қалытқыдай қалқуда... Мұның мәні – жай судың орнына салмағы ауыр су алынуында... «.

Бұл абзацты ***1-кесте*** үзіндісіндегі тұжырымдаудың төртінші санына жатқызуға болады, яғни тығыздығы 1 кг/м^3-ден асатын ***заттың*** (басты нысанның) суда бататындығын айырықша атап көрсету; басты субъектінің (ғалымның) басты нысанды (затты) суға батқызбай, оның су бетінде қалқып жүруі үшін жасалатын шаралары (судың тығыздығын 1 кт/м^3-ден де жоғары асыруы), болған жайтты қайталап атау (констатация фактов) ретінде қысқаша түрде баяндалуы.

Кестеде көрсетілгендей, формалды түрде осындай тұжырымдауға сай келетін абзац мазмұндарын «М004» деген шартты белгімен кодтау ұйғарылды. Кестедегі осы тәріздес шартты белгілер мәтін абзацтарының заттық-логикалық мазмұнын формалдау ісіне және солар арқылы мәтіннің қысқаша түрде мазмұнын ашу кезінде аса маңызды рөл атқарады деп санаймыз.

1-кесте

№ р/с	Кірісле абзацтар типтерінің тұжырымдалуы	Тип коды	Абсолютті жиілік (Fi)	Қатыстық жиілік (fi)
3	Басты нысанның бірінші және екінші дәрежелі нысандарға қарағанда өзіне ғана тән ерекшелігін қайталап атау (констатация фактов)	М003	11	0,1183
4	Ерекше көңіл аудару мақсатымен басты нысанды айырықша атап көрсету және оған байланысты автордың (тәжірибешінің) іс-әрекетін қысқаша кайталап атау	М004	13	0,1398
6	Кейбір нысандардың мәнін пайымдау (мақұлдау)	М006	10	0,1075
7	Автордың болашақ іс-әрекетіне байланысты шарттарды сөз ету	М007	10	0,1075
11	Автордың басты нысанға қатысты кейбір жай ғана пайымдауларын қайталап атау	М011	9	0,0968

Осылайша, ғылыми-көпшілікке арналған мәтіннің 353 медиалды (ішкі, ортаңғы) абзацтары да заттық-логикалық тұрғыдан талданып, соның негізінде 23 типтік тұжырымдаулар ажыратылды, Жиі қолданыс тапқан 5 түрлі тип ***2-кестеде*** көрсетілді.

Мәтін ішіндегі медиалды абзацтардың заттық-логикалық мазмұнын қарастыру барысында мынадай жайт байқалды. Мәтіннің ортаңғы абзацтарындағы басты субъекті мен автордың (тежірибешінің) кейбір іс-әрекеттерін (жүріс-тұрыс, бұлшық ет жұмысы мен қызметі) талдау арқылы басты нысанды таныту процесі іске асады.

Ғылыми көпшілікке арналған мәтін бойынша талданған медиалды абзацтың мысалы: *«Магнитті байытудың өнімділігін арттырудың бір жолы – магниттілігі аз минералдардың магниттік қасиетін алдын ала күшейту. Мысалы, гематит, сидерит және лимонит секілді минералдардың*

магниттілігі әлсіз. Егер осы минерал рудаларын 550-600 градуста көміртегін қосып күйдірсе күшті магниттеледі».

Заттық-логикалық мазмұны жағынан бұл медиалды (ортаңғы) абзацты ***2-кестедегі*** саны 10-шы тұжырымдауға «Басқа бір екінші дәрежелі нысандарға қатысты кейбір шарттарды қайталап атау» сай келеді деп есептеуге болады. Ал мәтіндегі мұндай тұжырымдауға бағынатын абзацтардың қолданылудағы абсолютті жиілігі Ғ=37, қатынастық жиілігі Ғ=0,1048 немесе олар орта есеппен алғанда мәтіндегі барлық абзацтардың 10,5 пайызын құрайды.

2-кесте

№ р/с	Медиалдыды абзацтар типінің тұжырымдалуы	Тип коды	Абсолютті жиілік (Ғі)	Қатыстық жиілік (fi)
10	Басқа бір екінші дәрежелі нысандарға байланысты кейбір шарттарды қайталап атау	М110	37	0,1048
11	Екінші дәрежелі нысанның ерекшелігіне арнайы көңіл аудару	М111	51	0,1445
13	Автордың өткен және осы шақтағы іс-әрекетін қайталап атау	М113	37	0,1048
16	Басты субъектіге қатысты аралық және қорытынды мінездемелердің қысқаша пайымдалуы	М116	41	0,1161
23	Автордың өткен шақтағы іс-әрекетін қайталап атау және басты субъектінің мінез-құлқын суреттеу	М123	23	0,0652

Ғылыми-көпшілікке арналған 47 мәтін ішінен алынған 57 қорытындылау абзацтарын зерттеу нәтижесінде олардың 10 түрлі типтік тұжырымдау топтары ажыратылды және статистикасы мен кодтау принциптері анықталды (***3-кесте***).

Ғылыми-көпшілікке арналған мәтіндер бойынша қорытынды абзацтарға қатысты тұжырымдаулардың кестеде көрсетілгендей мағынада берілуі өзінен-өзі түсінікті сияқты, себебі көп жағдайларда мәтін соңындағы абзацтарда болған іске қорытынды көзқарас айтылып, автордың болашақтағы іс-әрекетінің бағыт-бағдары белгіленеді.

Осы айтылғандарға мысал ретінде функционалды-мағыналық сипаты бойынша баяндау типіне жататын, ал заттық-логикалық мазмұн жағынан (біздің ұйғаруымызша) ***3-кестедегі*** рет саны 8 типтік тұжырымдауға жататын абзац мәтінін келтірейік:

«Атмосфера ауасының тазалығы жолында күрес жүргізу – ең алдымен адамдардың денсаулығын сақтау мәселесі. Сондықтан, адамдардың денсаулығына бірінші дәрежелі маңыз аударылатын елде – атмосфера ауасының гигиенасы гигиеналық ғылымның арнаулы саласы ретінде құрылып, ойдағыдай дамып келеді».

Бұл абзацтағы заттық-логикалық мазмұн бойынша басты нысан – ***атмосфера ауасы.*** Бірінші сөйлемде автор ***таза атмосфералық ауа*** мен ***адам баласының денсаулығы*** арасындағы тығыз байланыстың барына

ерекше көңіл бөле келе, басты нысанға, яғни ***атмосфералық ауаға*** сипаттама беруге әрекет етеді. Абзацтың екінші сөйлемінде бұл мәселені ғылыми тұрғыда қарастыру қажеттігі және оның біздің елде ойдағыдай дамып келе жатқаны сөз етіледі.

3-кесте

№ р/с	Қорытынды абзацтар типінің тұжырымдалуы	Тип коды	Абсол. жиілік (Fi)	Қатыс. жиілік (fi)
1	Басты нысанның мәні жайлы қорытынды нәтижелерді пайымдау	M201	19	0,3333
2	Басты нысанның ерекшеліктері жайлы автордың (зерттеушінің) пайымдауы	M202	10	0,1754
3	Басты нысанға қатысты автордың өткен шақтағы іс-әрекетін қайталап атау	M203	3	0,0526
4	Кейбір мәселелердің қойылуын кайта пайымдау және автордың болашақ іс-әрекетін сипаттау	M204	2	0,0351
5	Басты нысанға қатысты автордың болашақ іс-әрекетін қайталап атау	M205	1	0,0175
6	Басты субъектіге қатысты автордың қорытынды нәтижелерінің суретке (графикке, кестелерге және т.б.) байланысты пайымдалуы	M206	1	0,0175
7	Автордың алда тұрған іс-әрекеттерін қайталап атау	M207	1	0,0175
8	Мәселенің қойылуы мен оның бағалануы тұрғысынан басты нысанға берілген сипаттамалар	M208	13	0,2283
9	Басты нысанға қатысты екі жақты сипаттаманың қысқаша түрде берілуі	M209	4	0,0702
10	Басты нысан мен екінші дәрежелі нысандардың болашақ жағдаяты жайлы болжау айту	M210	3	0,0526
	Барлығы:		**57**	**1,0000**

Сонымен, осы зерттеу нәтижелері негізінде қазақ тіліндегі ғылыми-көпшілікке арналған мәтін мазмұнының формалды белгілер арқылы дұрыс ашылуы оның құрылымдық жағына әр уақытта тікелей байланысты бола бермейтіні және абзацтарды заттық-логикалық мазмұн тұрғысынан қарастырғанда осы стильге тән жалпы заңдылықтың сақталуымен бірге, ұқсас типтік тұжырымдаулар мәтіннің әр бөліктерінде де қолданыс табатыны айқындалды.

Жоғарыда сипатталған әдістер негізінде мазмұндық, типтерді ажырату жұмысы газет мәтіндері бойынша да жүргізілді. «Қазақстан мұғалімі» газетіндегі «Алты жасарлар мектепке барады» атты мәліметке қатысты (Атматы, 1986 ж.) жарияланған жалпы көлемі 502 абзац құрайтын тақырыптас 35 мақала компьютер көмегімен заттық-логикалық мазмұн тұрғысынан статистикалық жолмен зерттелді [1, 133-140].

Абзацтардың мәтін бойындағы орналасу тәртібі (***кіріспе, медиалды, қорытынды***) ескеріле отырып, олардың типтік тұжырымдаулары 3 түрлі

кесте түрінде көрініс тапты. Осы аталған зерттеу нәтижелеріне қысқаша тоқталайық.

Газет стилі бойынша 35 мәтін ішінен 25 бастапқы (***кіріспе***) абзац талданып, 11 мазмұндық типтің мәтін ішіндегі қолдану жиіліктері есептелді және әрбір типке қатысты шартты белгі-код қою принципі шешілді.

Газеттік стильъ бойынша 11 заттық-логикалық мазмұн типтерінің тұжырымдаулары ішінен ең жиі колданыс тапқандары мыналар:

– «басқа бір жанды (жансыз) нысандарға байланысты табиғатты немесе сырт жағдайды суреттеу»;

– «басқа бір нысанға сипаттама беру және басқа бір субъектінің болашақ іс-әрекетіне болжау айту»;

– «басқа бір нысанның мәні жайлы автордың пайымдауы және оның басты субъектіге қатысты қалпын қайталап атау»;

– «басты нысанға байланысты қайсыбір сырт жағдайды кайталап атау».

Газет стилі бойынша 175 медиалды абзац ішінен әр түрлі заттық-логикалық мазмұндағы 41 типтік тұжырымдаулар бөлініп алынды және олардың әрбіреуіне М401-М441 аралығын қамтитын шартты белгі-кодтар сәйкестікте қойылды. Бұл кодтар газет стилін формалды пішінде зерттеу кезінде қажеттік етеді. Мысал ретінде, типтік тұжырымдаулардың кейбір жиі қолданыс тапқандарын келтірейік:

– «басқа бір жанды (жансыз) нысандарға байланысты автордың кейбір деректерді кайталап атауы»;

– «басты субъектіге қатысты қайсыбір іс-әрекет жүргізілгенін қайталап атау»;

– «басты нысанға қатысты табиғатты немесе сырт жағдайды суреттеу»;

– «басты субъектінің іс-әрекетін қайталап атау және басты нысанға сипаттама беру»;

Мазмұн тұрғысынан заттық-логикалық, талдау газет мәтінінің 45 қорытынды абзацтарына да жүргізілді.

Зерттеу нәтижесінде типтік тұжырымдаудың 11 түрі топталып, қажетті санақ жұмыстары іске асты. Мысалы, мына тұжырымдауларды жеке бөліп атауға болады:

– «басты субъектіге қатысты автордың болжау айтуы»;

– «бірінші және екінші дәрежелі субъектілер (жанды, жансыз нысандар) жайлы автордың кең түрдегі пайымдаулары»;

– «мәтінде сөз болған оқиға жайлы автордың қорытынды пайымдау деректерін қайталап атауы».

Мысалы, *«Балабақшада оқиды»* атты мақала мынадай қорытынды абзацпен аяқталған: *«Алты жастағыларды оқытуда тек мектепке дейінгі мекемелерге арқа сүйеп жайбарақаттық танытпай мектептердің де мүмкіндіктерін молайту қажет. Бұл мәселе облыстық және қалалық халық ағарту бөлімдері басшыларынан барынша жауапкершілікті талап етеді».*

Бұл абзацтағы заттық-логикалық мазмұнға көңіл бөлсек, басты объектіден ***(мектепке дейінгі мекеме)*** басқа, мектептің өз мүмкіндігі, облыстық, қалалық халық ағарту бөлімдері мен олардың басшыларының

(жанды, жансыз нысандар) мүмкіншіліктері сөз болады. Сондықтан да, ең ықтималды типтік тұжырымдалу – «бірінші және екінші дәрежелі жанды және жансыз нысандар жайлы автордың кең түрде пайымдауы» типі сай келеді деп ұйғарған жөн сияқты.

Газет мәтінінің құрылымындағы барлық абзацтарды ғылыми-көпшілікке арналған стиль абзацтарымен салыстыра қарағанда, кіріспе, медиалды, қорытынды абзацтарды заттық-логикалық мазмұн тұрғысынан және құрылымдық жағынан қарастырғанда одардың өздеріне тән ерекшеліктерімен бірге ортақ тұстары да кездеседі. Сол сияқты ғылыми-көпшілікке арналған мәтін өзінің стереотиптік қасиетімен де ерекшеленетінін байқауға болады.

Жоғарыда сипатталған әдістер негізінде 20 түрлі көркем әдебиет мәтіндері заттық-логикалық мазмұн тұрғысынан типтік топтау процесімен өтті. Мұндай топтау жұмысы құрылымдық орналасу тәртібі ескерілмей алынған жалпы саны 487, мазмұны шағын оқиғаға негізделген абзацтардың бірінші, екінші және үшінші, т.б. дәрежелі жанды немесе жансыз нысандар рөлдерін ескеру арқылы іске асты [1, 140-143]. Осындай заттық-логикалық және статистикалық зерттеу жұмысының негізінде 487 абзац ішінен 32 типтік тұжырымдауды бөліп алуға мүмкіндік туды.

Төменде осы типтік тұжырымдаулардың тек жоғары жиіліктегілерін ғана келтірдік:

– «табиғатты немесе сырт жағдайды суреттеу және басты субъектінің іс-әрекетін қайталап атау» (М601, Ғ=54; f=0,1109);

– «табиғатты суреттеу және автордың іс-әрекетін немесе қалпын қайталап атау» (М604, Ғ=31; f=0,0636);

– «басты нысан мен басқа да басты субъектілерге қатысты негізгі басты субъектінің іс-әрекетін суреттеу» (М605, Ғ=40; f=0,08);

– «автордың болашақ іс-әрекетін қайталап атау» (М609, Ғ=40; f=0,0821);

– «авторға мінездеме беру және оның іс-әрекетін қайталап атау» (М624, Ғ=23; f=0,0472);

– «екінші дәрежелі субъектінің іс-әрекетін қайталап атау және авторға мінездеме беру» (М625, Ғ=20; f=0,0411);

– «басты субъектіні іс-әрекетке талаптандыру және автордың кейбір іс-қимылы мен қалпын қайталап атау» (М626, Ғ=31; f=0,0636);

– «басты субъектіні іс-әрекетке талаптандыру және басқа да басты іс-әрекетін қайталап атау» (М627, Ғ=20; f =0,0411).

Мысал ретінде, С.Сейфуллиннің ***«Бандыны қуған Хамит»*** атты әңгімесінен бір абзацты қарастырайық: *«Күн шайдай ашық, жылы еді. Аспанда бұлт жоқ, жел де жоқ. Айнала қалың иірім-иірім орман. Бойлап, таласа, жарыса шыққан қайың мен қарағай. Және әредік қалың қайың, қарағаймен жарыса шыққан теректер де көрінеді. Қайың мен терек ағаштардың жапырақтары салғырт тартып, сарғая бастаған. Осы орман ағаштардың ішімен, ескілеу жолмен Хамит жүріп келеді».*

Бұл абзац функционалды-мағыналық сипаты бойынша суреттеу абзацына жатады, құрылымдық жағынан – кіріспе абзацтардың бірі, ал заттық-логикалық мазмұн тұрғысынан алғанда тұжырымдау типінің М601 кодына жататын, қолданылу жиілігі бойынша барлық абзацтардың 11 пайызын құрайтын: «Табиғатты немесе сырт жағдайды суреттеу *(көркем орман іші)* және басты субъектінің іс-әрекетін *(Хамиттің орман ішімен жүруі)* қайталап атау» түріне жатады.

Көркем әдебиет стилі бойынша топталған 32 типтік тұжырымдауларды талдай келе байқағанымыз, олар көбіне табиғатты, қоршаған ортаны, сырт жағдайды суреттеу, бас кейіпкердің қалыпты жағдайын немесе іс-қимылын баяндау және басты субъекті мен объектіге мінездеме, сипаттама беру, автор мен басқа да субъектілердің өткен, осы және болашақтағы іс-әрекеттерін қайталап, болжай атау және т.б. арналады. Көркем әдебиет мәтіндерінің заттық-мағыналық тұрғыдан публицистика стилінен біршама, ал ғылыми мәтін стилінен айтарлықтай ерекшеленетіні байқалады.

Мақалада сөз болған зерттеу нәтижелері, яғни әр стиль бойынша ұсынылған типтік тұжырымдау топтары мәтін сипатына қарай сандық не сапалық жағынан өзгеріске ұшырауы мүмкін, біз тек бұл мәселенің принциптік жағын ғана қарастырдық деуге болады.

Сонымен, мәтін абзацтарын функционалды-мағыналық мазмұнға топтау болсын, заттық-логикалық мазмұн тұрғысынан типтік тұжырымдау топтарына ажыратуға арналған зерттеулер болсын, бәрінің мақсаты – қазақ тіліндегі мәтін мазмұнын формалды тұрғыдан ашу қажеттілігінен туындайды.

Әдебиет

1. *Жубанов А.К.* Основные принципы формализации содержания казахского текста. Алматы, 2002. 250 с.
2. *Зубов А.В.* Проблемы порождения текста. Ч. II. Вероятностно-алгоритмическая модель порождения текста. Минск, 1989. 63 с.
3. *Гришина О.Н.* Проблемы контекстно-вариативного членента текста в стиле языка художественной и научной прозы // Функциональные стили и преподавание иностранных языков. М.: Наука, 1982. С. 52-68.

ҚҰДАЙБЕРГЕН ЖҰБАНОВ ЖӘНЕ ҚАЗАҚ ТІЛ БІЛІМІНДЕГІ СТАТИСТИКАЛЫҚ ӘДІСТІҢ ЖЕТІСТІКТЕРІ

ҚР БҒМ Ақтөбе мемлекеттік педагогикалық институты
Жұбанов тағлымы: VI Халықаралық ғылыми-теориялық конференция материалдары. –Ақтөбе, 2005. 24-29-беттер.

Статистикалық әдіс басқа математикалық әдістерге қарағанда тіл зерттеу ісіне бұрынырақ және тұрақты сипатта енді деуге болады. Бұл әдістің

қажеттілігі айқын, ол қазіргі жағдайда тілшілер арасында айтарлықтай күмән туғызбайды.

Белгілі орыс ғалымы Б.Н.Головин әуелден-ақ сандық сипаттың тіл категориялары мен оның элементтеріне объективті түрде тән болуынан, сонымен бірге сапалық және сандық сипаттарының өзара катынастылығы тілдің қатынас құралы ретіндегі қызметі мен даму сатысындағы объективтілігінен тілді сандық тұрғыда зерттеу қажеттілігі туындады деп түсіндіреді [1]. Мысалы, морфемдік құрам сөздің сапасына әсер ететіндігі сияқты морфологиялық деңгейдегі сандық құбылыстар оның синтаксистік деңгейдегі сапасына әсер етпей қоймайды. Осыдан барып әр қалам иесінің стилі де тек өзіне ғана тән тұрақты бірліктерінің қатынастары арқылы ерекшеленеді.

Заңдылықтарды тәжірибелік байқау мен оның нәтижелерін математикалық жолмен өңдеу арқылы анықтау тілдік жүйені танып-білудегі лингвистикалық статистикадағы әрі табиғи, әрі ғылыми бетбұрыс болып саналады. Нәтиженің дәлдігіне, жасалатын тұжырымдардың объективті болуына талпыну – барлық ғылымдар саласына тән жағдай. Ал қоғамдық ғылымдар саласындағы ерекше дәлдікті қажет ететін ғылым саласы лингвистика деуге болады.

Статистикалық әдістің тіл саласындағы колданылу тарихын сөз еткенде, осы әдісті тұңғыш ұсынған орыстың белгілі математик-ғалымы В.Я.Буняковскийдің (1804-1889) атын атамай кетуге болмайды [2]. Сонымен бірге басқа да ірі ғалымдардың тілді зерттеуде сандық деректерге жүгінудің кажеттілігін дәлелдеулеріне деректерді көптеп келтіруге болады (И.А.Бодуэн де Куртенэ, А.М.Пешковский, М.Н.Петерсон, Е.Д.Поливанов, В.В.Виноградов және т.б.).

Классикалық тіл білімінің өкілі В.В.Виноградов тілдің кітаби, сөйлеу түрлерінде және көркем әдебиет стилінің әр жанрларында сөздердің қолдану жиілігінің біркелкі еместігін айта келе, оларды анықтау үшін арнайы зерттеудің кажеттігін көрсетеді. Ғалымның тұжырымдауынша, мұндай зерттеулер түрлі стильдердің құрылымдық-грамматикалық, тіпті семантикалық айырмашылықтарын белгілеуге, сондай-ақ барлық грамматикалық категорияларына статистикалық талдау жүргізілген әр әдеби стильдің қатынастық-функционалдық салмағын айқындауға мүмкіндік туғызады [3].

Мәселен, Ф.П.Филин өзінің бір еңбегінде: «Ғылымның бір-ақ мақсаты болады, ол – әлі ашылмаған шындықты ашу немесе соның ашылуына ықпал жасау, сол арқылы қоғамға қызмет ету. Алға койған мұндай мақсатқа әр түрлі жолмен жетуге болады. Соның ішінде зерттеудің дәлдігі мен объективтілігін қамтамасыз ететін тәсіл ғана ең дұрысы болмақ. Дәлдік пен объективтілікке ұмтылу – қазіргі ғылымның туы. Бұл тіл біліміне де қатысты», – деп жазса [4], В.Н.Головин тіл ғылымына сандық сипаттың тән екендігін, сондықтан тілші-ғалымдардың көпшілігінің тілдік деректерді сипаттағанда сандық ұғымдарға жүгінетінін айтады [1].

Тілдік фактілерді сандық ұғымдарға жүгініп баяндау қазақ тіл білімін зерттеуші ғалымдарға да тән. Өйткені қай кездегі болмасын қазақ тілі жайлы зерттеулердің бәрінде дерлік ***аз, көп, көптеген, жиі, сирек, өнімді, өнімсіз, мол, жуық, тең, шамалас, көп рет*** т.б. шама, мөлшерді көрсететін сөздер кездесіп отырады және бұлардың мәнін нақты санмен көрсетуге талаптанушы ғалымдар да бар. Сөйтіп, қолданбалы тіл білімінің бүгінде «Статистикалық лингвистика» деп аталып жүрген жаңа саласының алғашқы нышандары мен белгілері міне осылардан басталады. Ал қазақ тіл біліміндегі бұл саланың бастауында профессор Құдайберген Қуанұлы Жұбанов тұрды. Мұны белгілі тілші-статист ғалымдар – К.Б.Бекбаев, С.Мырзабеков т.б. қазақ тілінің статистикасы туралы зерттеулері мен мақалаларында атап көрсеткен. Мәселен, С.Мырзабеков «Қазақ тілін зерттеуде санды пайдалану» атты мақаласында: «Қазақ совет тіл білімінің негізін қалаушылардың бірі, совет дәуірінде қазақ тілін тұңғыш рет ғылыми түрде зерттеген маман-лингвист Қ.Жұбанов ретіне қарай сандық деректерді де пайдаланған», – десе [4, 167], ғалым Қ.Б.Бектаев «Лингвистикалық статистикаға катысты элементтерді – сандық деректер мен мәліметтерді, оны пайдалану нәтижесіндегі кейбір пікірлер мен тұжырымдарды біз профессор Қ.Жұбановтың еңбектерінен жиі кездестіреміз», – дейді [5, 110]. Ғалымның сол кездің өзінде-ақ тіл білімінде сандық деректерді пайдаланудың қажеттілігіне ерекше ден қойғаны байқалады. Мәселен, ғалымның 1936 жылы жарық көрген «Қазақ тілінің грамматикасында» сөз мүшелерінің орны, дыбыстардың саны мен сапасы, түрлері туралы жазған мына тұжырымы көңілге қонымды: «Дыбыстың саны бар да, сапасы бар. Дыбыстың сапасы бірдей бола тұрып, саны әлденешеу бола береді. ...Бір тілде жұмсалатын дыбыстардың жалпы санының ұшы-қиыры жоқ, бірақ оның бәрінің бірдей сапасы әр түрлі бола бермейді. Сапасы әр түрлі болатын дыбыстар да, көп болмайды». Сол сияқты «... Ілиястың небәрі 12 жол «Қойшы ойы» деген кішкене өлеңінде 488 дыбыс бар. Бірақ 488 дыбыс түрлі емес, мұнда небәры 28-ақ түрлі дыбыс бар. Осы 28 түрлі дыбыстың бірі 10; бірі 50 рет қайталанып барып 488 болған. ... Қазақ тілінде де сапасы әр түрлі болатын дыбыстардың арнаулы саны бар» [6, 183].

Профессор Қ.Қ.Жұбановтың «Қазақ тілі жөніндегі зерттеулерінде» бұл жайлардан басқа да мәселелер, айталық: сөз тіркесінің формалары, сөйлем мүшелерінің орын тәртібі, буындар және ондағы дыбыс саны, буындарды оқыту әдістемесі т.б. мәселелер сөз болған тұста көптеген сандық (статистикалық) деректер мен мәліметтер келтірілген [6].

Шын мәніндегі статистикалық ізденістерді Қ.Жұбанов қазақ орфографиясын ғылыми негізде құру мақсатында қолданған. Ол әліпби құрамына енуге тиісті әріптердің санын қазақ тіліндегі белгілі бір фонеманы іс жүзіндегі қажеттігіне, оған деген ділгерлікке, қолданыс жиілігіне бағындыра белгілеуді мақсат тұтқан [5, 114].

Міне, бұл айтылғандардан проф. Қ.Қ.Жұбановтың тілді зерттеуде, оның заңдылықтарын қарауда есеп-санаққа, ретті жерінде математикалық ойлауға, оның тәсілдеріне сүйенгендігінің тағы да бір дәлелі.

Қ.Қ.Жұбанов енбектеріндегі сандық деректерді пайдалану жайымен таныса отырып, мынадай қорытынды жасауға болады: проф. Құдайберген Қуанұлы Жұбановтың қазақ тіл білімі туралы зерттеулерінде тілді сандық деректер мен мәліметтер негізінде қараудың алғашқы нышандары бар. Бүгінде тіл білімінің бір саласына айналып отырған лингвистикалық статистиканы ғалым сол кездің өзінде-ақ тани білген. Тілдің сандық сипатынан байқалатын сапалық қасиеттерін оны зерттеуде пайдалануға болатын тиімді әдіс көздерінің бірі екенін де байқаған. Сөйтіп, ол сандық деректердің негізінде тілге қатысты бірсыпыра жайларды (дыбыстар, буындар, сөздер, сөйлем мүшелерінің орны т.б.) нақты анықтап, кейбір тың тұжырымдар жасаған. Бұлардың маңыздылығы сонда – олар қазіргі кезде бұл саладағы ізденуші-зерттеушілерге үлгі-тірек, бағыт-бағдар болып отыр [5,115-116].

Қазақ тілінің нормативті курсын жүйелеу жұмысына проф. Қ.Жұбановтан кейін белгілі тілші-ғалымдар: Н.Т.Сауранбаев, С.К.Кеңесбаев, М.Б.Балақаев, А.Ы.Ысқақов және т.б. ғалымдар атсалысты. Бастауыш және жоғары сыныптарға және педучилищелерге арналған қазақ тілі оқулықтарының авторлары (қалам иелері) тілдің буындық құрылымын, дыбыс физиологиясын, тілдің акценттік және екпіндік және т.б. мәселелеріне қатысты зерттеулерінде сан мен сапаны бөлмей, бірге қарастырады. Мәселен, С.К.Кеңесбаев қазақ тіліндегі буын типтерінің, М.Б.Балақаев қазақ тіліндегі синтаксистік сөз тіркестерінің табиғатын зерттеуде көптеген нақты мәтіндік материалдарды сандық қатынастар көмегімен қарастырған [7; 8]. Арнайы сандық талдауларды профессор Ғ.Мұсабаевтың зерттеулерінен де кездестіруге болады [9]. Әрине, ғалымдардың мұндай ізденістері, яғни сандық қатынасқа көңіл бөлуі саналы түрдегі ықыластан гөрі тілдің ішкі табиғатына (сырына) катысты туындаған деп пайымдауға болады.

Осы кезеңде қазақ тілінің грамматикалық форманттарын статистикалық жолмен зерттеуге айрықша көңіл бөліне бастаған болатын. Қазақ лексикографтары сандық деректі ұлы ақын Абай тілінің сөздігін түзу кезінде жүйелі түрде толық пайдаланды. Тұңғыш рет қазақ тіл білімінің тарихында бір ғана автордың (Абайдың) жазба түріндегі шығармалар тілі толық қамтылып, әр сөзі мен сөз тіркесінің қолдану жиілігіне көңіл бөлінді. Яғни Абайдың тіл байлығы, саны мен сапасы жайлы жайттар тілшілердің арнайы зерттеу нысанына айналды. Сөйтіп, орыстың ұлы ақыны Пушкин тілі сөздігі тәріздес «Абай тілі сөздігі» жасалды [10]. Бұл сөздікті әрі түсіндірме сөздік, әрі әліпби-жиілік сөздік деуге болады. Сөздердің қолдану жиілігін көрсету сөздік түзудің негізгі мақсаттарының бірі болып саналды.

1969 жылы Қазақстан Ғылым академиясы Тіл білімі институтында «Статистическое и информационное изучение тюркских языков» атты Бүкілодақтық ғылыми жиынның өтуі қазақ тіл білімі үшін елеулі құбылыс болды. Осы ғылыми форумның нәтижесі ретінде 1970 жылы Тіл білімі институтында статолингвистика саласының белгілі ғалымы Қалдыбай Бектаевтың жетекшілік етуімен «Тіл статистикасы және автоматтандыру» атты ғылыми топ құрылды. Осы кезеңнен бастап қазақ тілін статистикалық

зерттеу ісі жүйелі сипатқа ауысты. Бірнеше жас ғалым статистикалық лингвистика саласынан кандидаттық диссертация қорғап, олардың ғылыми мақалалар жинағы, монографиялары, оқу құралдары жарық көрді. Кеңес Одағы кеңістігіндегі ғылыми конференцияларға кеңінен жол ашылды. Қазақстан статолингвистер зерттеулерінің нәтижесі ретінде 1973 жылы «Қазақ тексінің статистикасы» атты ғылыми жинақ жарық көрді [11]. Бұл жинақ сол жылға дейінгі қорғалған (және дайындалған) кандидаттық диссертациялардың негізгі мазмұнын ашады деуге болады.

Тілді зерттеу тәжірибесінде әр тілді жалпы типологиялық тұрғыдан бағалауда қолданылатын статистикалық критерийлердің үлес салмағын «өлшеуге» (анықтауға) кажетті жалпылама сандық дәреже әлі де ғалымдарды толық қанағаттандырмай келеді. Міне, осы бос кеңістікті толтыруға арналған іргелі ғылыми зерттеу ретінде 1978 жылы басылып шыққан Қ.Б.Бектаевтың «Статистико-информационная типология тюркского текста» атты монографиясын атауға болады [12]. Бұл зерттеу түркі мәтіндері үшін ықтималдық және ақпараттық үлгілеу әдістерін жүйелі түрде қолдануға арналды. Ғалым түркі тілдерінің типологиялық ерекшеліктерін үндіеуропа тілдерімен салыстыра зерттеуді мақсат етіп қойып, оны формалды статистикалық-ықтималдық әдіс арқылы іске асырды.

Осы кезеңдік зерттеулерге тән мәселелер – ол тілдің фонетика мен фонология саласы, сөз бен сөзтұлғаның буындық құрылымы, қазақ мәтінінің түрлі функционалды стилінің лексика-морфологиялық құрылымы, қазақ тілінің ақпараттық сипаты және т.б. кешенді проблемалар болатын. Негізгі зерттеуге қажетті статистикалық құрал ретінде электронды-есептеу машиналары (ЭЕМ) көмегімен жасалатын әліпби-жиілік сөздік, жиілік сөздік және кері әліпби жиілік сөздіктер материалдары болды. Жиілік сөздіктерді ЭЕМ көмегімен автоматты түрде түзу мәселесі мен олардың құнды деректерін тілді зерттеу практикасында іске асыру мүмкіндігі жайлы Кеңес Одағы кеңістігі ғылыми орталықтарында ұйымдастырылған көптеген Бүкілодақтық ғылыми жиындар болып өтті. Қазақ тілін статистикалық әдіспен зерттеу және оны автоматтандыру мәселелері де Одақтық дәрежеде қарастырылды. Отандық түркітану ғылымы үшін ортағасырлық жазба ескерткіші «Кодекс Куманикус» мәтіні тұңғыш рет ЭЕМ жадына енгізіліп, ол бойынша автоматты түрде «Куманша-қазақша жиілік сөздігінің» алынуы қазақ тілі тарихы үшін елеулі оқиға деуге болады [13].

Жоғарыда сөз болған қазақ тілі мәтіндерін статистика тәсілімен зерттеуге арналған кандидаттық диссертациялар тақырыбы 1979 жылы «Ғылым» баспасынан жарық көрген М.О.Әуезовтің «Абай жолы» романының жиілік сөздіктеріне тікелей не салыстырмалы тұрғыда қатысты болды [14]. Ал ЭЕМ көмегімен алынған және 1995 жылы басылып шыққан «М.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері» ұлы жазушының 100 жылдық мерейтойына арналған Қазақстандық статолингвистерінің арнайы сыйы деуге болады [15]. Жазушы тілінің лексикалық байлығын, автордың сөз қолданудағы стильдік тәсілінің өзіндік ерекшелігін, қазақ тілінің көркемдік және бейнелеуші мүмкіндіктерін

меңгеру дәрежесін осы аталған лексикографиялық еңбек бойынша бағалауға болады. Осымен қатар бұл сөздік жалпы қазақ лексикасын статистикалық құрылым тұрғысынан зерттеуге бай тілдік материал бола алатыны сөзсіз. Кітап түрінде баспадан жарық көрмегенімен М.Әуезовтің әр томы бойынша компьютер жадында сақталған «Сөзнұсқағыш әліпби жиілік сөздік» атты электронды сөздік базасы да арнайы қаралатын лексикографиялық еңбек деуге түрады.

А.Ахабаевтың қазақ тілін нормативті тұрғыда тануға, қазіргі қазақ публицистикасы тілінің лексика-морфологиялық құрылымын статистикалық талдауға арналған ғылыми жұмысы да маңызды зерттеу болып табылмақ [16]. Бұл жұмыстың шағын бөлігі дублет сөздер вариантының нормасын статистика жолымен айқындау мәселесіне арналган. Статистикалық талдау негізінде А.Ахабаев сөздердің жиілік сипаты негізгі объективті критерий болатындығын сөз етеді. Жалпы алғанда, бұл ғылыми жүмыс 1965-1966 жж. газет мәтіндері мен «Абай жолы» романы (2-ші кітап) мәтініндегі зат есім мен есімдік жалғауларының форма-варианттарын статистика-морфологиялық тұрғыда салыстыра талдауға арналып, соның негізінде қазақ тілінің грамматикалық тұлғаларының нұсқалары стильдік айырым белгілер болатыны анықталған.

Статистика-морфологиялық зерттеудің келесі бір нысаны – қазіргі қазақ тілі етістік сөздерінің туынды тұлғаларын құрайтын негіз етістік пен оларға жалғанатын аффикстер. Осы зерттеуді М.Әуезовтің «Абай жолы» романындағы туынды түбір етістіктерінің құрылымына статистикалык талдау жасау арқылы ғалым С. Мырзабеков жүргізді. Ғалымның сандық деректері бойынша романда кездесетін барлық түбір етістіктің 80 пайызы – туынды етістіктер екені және етістіктің туынды түбірлері негізгілерден төрт есе көп қолданылатыны анықталды [17]. Етістік құрылымына жасалған лингвостатистикалық талдаудың нәтижелері қазіргі қазақ тіліндегі етістік сөздерді танып-білуге мүмкіндік беретіні сөзсіз.

Тіл ерекшелігі мәтіндердің жанрлық (стильдік) түрлеріне қарай да айырым табатыны белгілі. Осындай айырымдарды қазақ тілінің әр түрлі жанрларына қатысты статистикалық жолмен тағайындауды қазақстандық статистика тобының ғалымдары 70-80-жылдары өз мақсаты етіп қойды.

Қазақ тілінің фонетика саласы да статистика әдісінің назарынан тыс қалған жоқ. Фонетиканың дәстүрлі (есту) және құралды-тәжірибелік зерттеу әдісі бойынша фонологиялық және графикалық жүйелерге қатысты мәселелерін ыктималды-статистикалық әдіс-тәсілдермен шешу тиімді деп танылды.

Қазақ тіл білімінің қолданбалы саласының тың проблемалық мәселесі – қазақ мәтінінің әр әріпке (дыбысқа) шаққандағы ақпараттық өлшемін, яғни ***энтропияны*** анықтау. Осындай зерттеулер белгілі ғалым Қ.Б.Бектаев пен оның шәкірті Д.А.Байтанаеваның қазақ мәтініне қатысты ақпараттық-статистикалық ғылыми ізденістері арқылы іске асырыла бастады. Табиғи тілдің белгілер жүйесі үш түрлі деңгейде – ***синтактикалық, семантикалық*** және ***прагматикалық*** аспектілерде қарастырылатыны мәлім. Мұндағы

синтактика деңгейі әр түрлі белгілер жүйесінің синтаксисін, яғни сөз мағынасынан тыс, белгілердің тіркесімдік құрылымын және олардың туындау ережелерімен қатар белгілер құрылымының ішкі қасиеттерін қарастырады. Ал белгілер жүйесінің сыртқы құрылымдық қасиеттерімен семантика мен прагматика айналысады. Дәлірек айтсақ, егер ***семантика*** белгілер жүйесін мағыналық көзқарас, яғни мәтін мазмұны тұрғысынан танып-білумен шұғылданса, ***прагматика*** белгілер жүйесінің қабылдаушыға қатынасын анықтау мәселелерімен айналысады.

Қазіргі кезде ақпарат аясын теориялық және технологиялық тұрғыда зерттеудің синтактикалық денгейі жақсы дамып келеді. Бұған К.Шеннонның символдар жүйелілігі мен комбинаторикасын қарастыратын статистикалық байланыс теориясы негіз болады. Бұл теория бойынша табиғи тілде жазылған мәтін белгілі бір код құрайтын дискретті белгілердің тізбегі ретінде қарастырылады. Ал мұндай мәтінді қабылдау кезінде дискретті белгілердің қай-қайсысының да пайда болуы белгісіздіктің шамасымен, яғни энтропиямен өлшенеді. Қ.Б.Бектаев пен Д.А.Байтанаеваның ғылыми ізденістерінде қазақ мәтінінің ең кіші бірлігі – әріпке түсетін энтропияның орта шамасын және қазақ тілінің толық және стильдік (жанрлық) айырымдарын анықтайтын жазба мәтіндеріндегі энтропия мен артық ақпарат шамаларын анықтау негізгі мақсат ретінде қойылды. Қазақ мәтініндегі әрбір әріпке түсетін энтропияның ең үлкен (максимум) шамасын анықтау, әліпби әріптерін қабылдаудың ықтималдықтары тең және өзара тәуелсіз дәрежеде деген ұйғарымға негізделетіні айкын [8].

1987 жылы Қаз. КСР «Наука» баспасынан шыққан (осы жолдар авторының) «Квантитатавная структура казахского текста» және 2002 ж. жарық көрген «Основные принципы формализации содержания казахского текста» атты монографияларында қазақ тілінің заманға сай өзекті зерттеу мәселелері сөз етілді. Бірінші аталған ғылыми еңбекте қазақ тіліндегі мәтіндерді автоматты түрде электронды-есептеу машиналары (ЭЕМ) арқылы статистикалық әдіспен зерттеу мүмкіндігі мен сөз таптарының ықтималды-үлестірімді заңдылықтары зерттелсе, екінші монографияда – қазақ мәтінінің мәтін лингвистикасы тұрғысынан мазмұнын формалдау жолдары және мәтіннің туындау (порождение) проблемалық мәселелері қарастырылады [19; 20].

Сонымен, Құдайберген Жұбанов зертеулерінде тұңғыш рет орын алған және одан кейінгі ғалымдардың ғылыми ізденістерінде жалғасын тапқан статистикалық әдіс жеке автор шығармалары, функционалдық стильдер не толық әдеби тіл қамтылған мәтіндер бойынша зерттеу тәжірибесінде әр түрлі жиілік сөздіктер жасаудың өзекті мәселе екендігін көрсетті. Себебі олардың негізінде аса маңызды және күрделі мәселелерді: тілді лексика-грамматикалық нормалау, түркі тілдерін өзара салыстырмалы-типологиялық зерттеу және олардың компьютерлік қорын жасау мәселелерін шешуге зор мүмкіндік жасалады. Сол сияқты, қазақ тілінің көптеген кезек күттірмейтін мәселелері, мәселен, лексикография саласы бойынша компьютерлік қор (база) жасау, қазақ тілінен басқа тілге, не керісінше компьютерлік аударма

жасау, ақпарат мазмұнынан қысқаша реферат, аннотация алу және қолданбалы лингвистика саласының басқа да жайттары өз шешімін күтуде.

Әдебиет:

1. *Головин Б.Н.* Язык и статистика. М., 1971. 191 с.

2. *Струве П.Б.* Кто первый указал на применение статистики к филологическим исследованиям? // Изв. Российской АН. Сер. 6. 1918. №13.

3. *Виноградов В.В.* Современный русский язык. Введение в грамматическое учение о слове. 1938. Вып. 1. 160 с.

4. *Мырзабеков С.* Қазақ тілін зерттеуде санды пайдалану // Қазақ тілі грамматикасы бойынша зерттеулер. Алматы, 1975.

5. *Бектаев Қ., Белботаев А., Смайылов Т.* Қазақ тіл біліміндегі лингвистиканың алғашқы көріністері // Құдайберген Жұбанов және қазақ совет тіл білімі. Алматы: «Ғылым», 1990. 108-115-бб.

6. *Жұбанов Қ.* Қазақ тілі жөніндегі зерттеулер. Алматы: «Ғылым», 1999. 581 б.

7. *Балакаев М.* Современный казахский язык. Алма-Ата, 1985.

8. *Кеңесбаев І., Мұсабаев Ғ.* Қазіргі қазақ тілі. Алматы, 1962. 315-б.

9. *Мұсабаев Ғ.* Қазіргі қазақ тіліндегі сын есімнің шырайлары. Алматы, 1951. 90 б.

10. Абай тілі сөздігі. Алматы, 1968. 734-б.

11. Қазақ тексінің статистикасы. Алматы, 1973.731-б.

12. *Бектаев К.Б.* Статистико-информационная типология тюркского текста. Алма-Ата, 1978. 183 с.

13. *Құрышжанов А.Қ., Жұбанов А.Қ., Белботаев А.Б.* Куманша-қазақша жиілік сөздік. Алматы, Ғылым, 1978. 277 с.

14. *Бектаев Қ.Б., Жұбанов А.Қ., Мырзабеков С., Белботаев А.Б.* М.Әуезовтің «Абай жолы» романының жиілік сөздігі. Алматы, 1979. 334-б.

15. *Бектаев Қ.Б., Жұбанов А.Қ., Мырзабеков С., Белботаев А.Б.* М.О.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері. Алматы-Түркістан, 1995. 346-б.

16. *Ахабаев А.* Статистический анализ лексико-морфологической структуры языка казахской публшдистики. Автореф. дис. ... канд. филол. наук. Алма-Ата, 1971. С. 24.

17. *Мырзабеков С.* Статистико-лингвистический анализ структуры глагола современного казахского языка. Автореф. дис. ...канд. филол. наук. Алма-Ата, 1973. С. 32.

18. *Байтанаева Д.А., Бектаев К.Б.* Энтропия казахского текста // Статистика казахского текста. Алма-Ата, 1973. С. 664-696.

19. *Джубанов А.Х.* Квантитативная структура казахского текста (опыт лингвистического анализа на ЭВМ). Алма-Ата, 1987. 147 с.

20. *Жубанов А.К.* Основные принципы формализации содержания казахского текста. Алматы, 2002. 250 с.

ҚАЗАҚ ТІЛТАНЫМЫНДАҒЫ СТАТИСТИКАЛЫҚ ӘДІСТІҢ ОРНЫ

Аль-Фараби атындағы ҚҰУ. Э.Д.Сүлейменованың мерейтойына арналған «Жизнь языка и язык в жизни» атты жинақ. Алматы – 2005. 115-129-бб.

Өткен ғасырдың өзінде үнді-еуропа тіл білімінде кең тараған статистикалық әдіс түркітануда, соның ішінде қазақ тілінде тек XX ғасырдың 60-жылдарынан бастап жүйелі және ғылыми тұрғыда қолданыла бастады. Бірақ бұдан бұрын да аталған әдіс ішінара колданыс тауып, тілдің әр деңгейіндегі бірліктердің сандық және пайыздық ара-салмағын анықтаумен шектеліп келген болатын.

Түркі тіліндегі мәтіндерді статистикалық тәсілмен зерттеу – бұл тілдің ішкі табиғатына, яғни агглютинативті құрылымына негізделеді. Мамандардың пайымдауынша, кітаби және сөйлеу тіліндегі түрлі типті сөздердің әр ыңғайдағы қолдану жиілігін анықтайтын мұндай тәсіл тілдің құрылымдық-грамматикалық, кейде стильдік-семантикалық айырым белгілерін ажырата отырып, осы стильдерге тән қасиеттерді анықтауға мүмкіндік жасайды. Сонымен бірге грамматикалық категорияларды сандық тәсілмен талдау олардың әдеби тілдің барлық аясындағы функционалдық салмағын көрсетуде де маңызды.

Белгілі филолог-статист Р.Г.Пиотровский түркі тілдерін сандық деңгейде зерттеудің негізі олардың ішкі құрылымдық табиғатында жатқандығын айта келе, бұл тілдерге үнді-еуропа тіліне тән екпінді ассимиляция мен редукциялар және осы сипатты қасиеттер жат екенін баса көрсетеді. Бұл жағдай тілдік деректердің жай кезде қиындық туғызатын кейбір тұстарын математикалық әдіспен тіркеу мен тануда аса маңызды деп саналады [Пиотровский 1969].

Тілдік мәселелерге қолданылатын математикалық әдісті интуитивті түрде тұжырымдалған, сол сияқты толық шешімі жоқ мәселелердің логикалық тұжырымдалуына бағынатын және алгоритмдік шешімі бар бір немесе бірнеше қарапайым математикалық есептермен ауыстыру деп ұғыну қажет. Тілдік мәселені осылайша математикалық жолмен шешу – лингвистикалық нысанның математикалық моделіне (үлгісіне) көшуді білдіреді [Джубанов 1987: 6]. Тілдің аса күрделі түрдегі сан алуан параметрлі көп қабатты жүйе екенін ескерсек, онда ықтималдық теория, математикалық статистика, ақпарат теориясы және т.б. ғылыми сала әдістерін қолданудың мүмкіндігі кең екені байқалады. Әсіресе, бұл әдістер сөйлеу тәжірибесінде (жазба не ауызша) жиі кездесетін белгілі бір кешенді шарттарға байланысты тілдік кездейсоқтық құбылыс заңдылықтарын қарастырғанда аса ыңғайлы. Тіл табиғатының нақты ғылымдар саласының ережелеріне бағына бермейтін тұстары баршылық. Сондықтан осындай күрделі нысан ретіндегі тіл саласына ықтималдық теориясы мен математикалық статистиканы қолдану үшін, ең алдымен, «тілге» бірнеше шектеулер қою қажет. Мәселен, толығымен

өзгермелі бірліктерден тұратын табиғатына жүйелілік сипат тән тілді көптеген зерттеуші «ашық жүйе» ретінде қарастыруды дұрыс көреді. Осы тұжырымның негізінде, тілге қойылатын бірінші шектеу бойынша, белгілі-бір кезеңде тілдік жүйеде өзгермелілік сипат болмайды деп, оны «жабық жүйе» ретінде санау (модельдеу) ұйғарылды.

Екінші шектеу – тілдің ауызша және жазбаша түрлерінің аралық айырым белгілерін ескермей, оны арнайы түрде құрастырылған белгілер жүйесі немесе жазба тіл деп қарастыру. Мұндай ұйғарым мәтінді көру арқылы қабылдауға негізделген сипаттау грамматикасын құруға жеткілікті негіз бола алады [Пиотровская 1911].

Тілге қойылатын келесі шектеудің туындау себебі: тілдің қай саласында болмасын зерттеу арқылы сөйлеу қызметінің әр алуандығын қамту мүмкін еместігінде. Өйткені зерттеу нысаны ретінде алынған мәтінге қатысты мәліметтер тек сол мәтіндік стильге ғана тән болады. Зерттеу барысында мәтін типі мен оның сандық мөлшерін дұрыс таңдай білу зерттеу арқылы байқалған заңдылықтың дұрыстығына күмән туғызбаудың айғағы деп саналады.

Статистикалық әдіс басқа математикалық әдістерге қарағанда тіл зерттеу ісіне бұрынырақ және тұрақты сипатта енді деуге болады. Бұл әдістің қажеттілігі айқын, ол қазіргі жағдайда тілшілер арасында айтарлықтай күмән туғызбайды.

Белгілі орыс ғалымы Б.Н.Головин әуелден-ақ сандық сипаттың тіл категориялары мен оның элементтеріне объективті түрде тән болуынан, сонымен бірге сапалық және сандық сипаттарының өзара қатынастылығы тілдің қатынас құралы ретіндегі қызметі мен даму сатысындағы объективтілігінен тілді сандық тұрғыда зерттеу қажеттілігі туындады деп түсіндіреді [Головин 1971]. Мысалы, морфемдік құрам сөздің сапасына әсер ететіндігі сияқты морфологиялық деңгейдегі сандық құбылыстар оның синтаксистік деңгейдегі сапасына әсер етпей қоймайды. Осыдан барып әр қалам иесінің стилі де тек өзіне ғана тән тұрақты бірліктерінің қатынастары арқылы ерекшеленеді.

Статистикалық әдістің тіл саласындағы қолданылу тарихын сөз еткенде, осы әдісті тұңғыш ұсынған орыстың белгілі математик-ғалымы В.Я.Буняковскийдің (1804-1889) атын атамай кетуге болмайды [Струве 1918]. Сонымен бірге басқа да ірі ғалымдардың тілді зерттеуде сандық деректерге жүгінудің қажеттілігін дәлелдеулеріне деректерді көптеп келтіруге болады (И.А.Бодуэн де Куртенэ, А.М.Пешковский, М.Н.Петерсон, Е.Д.Поливанов, В.В.Виноградов және т.б.).

Классикалық тіл білімінің өкілі В.В.Виноградов тілдің кітаби, сөйлеу түрлерінде және көркем әдебиет стилінің әр жанрларында сөздердің қолдану жиілігінің біркелкі еместігін айта келе, оларды анықтау үшін арнайы зерттеудің қажеттігін көрсетеді. Ғалымның тұжырымдауынша, мұндай зерттеулер түрлі стильдердің құрылымдық-грамматикалық, тіпті семантикалық айырмашылықтарын белгілеуге, сондай-ақ барлық грамматикалық категорияларына статистикалық талдау жүргізілген әр әдеби

стильдің қатынастық-функционалдық салмағын айқындауға мүмкіндік туғызады [Виноградов 1938].

Тілдік фактілерді сандық ұғымдарға жүгініп баяндау – қазақ тіл білімін зерттеуші ғалымдарға да тән. Өйткені қай кездегі болмасын қазақ тілі жайлы зерттеулердің бәрінде дерлік *аз, көп, көптеген, жиі, сирек, өнімді, өнімсіз, мол, жуық, тең, шамалас, көп рет* т.б. шама, мөлшерді көрсететін сөздер кездесіп отырады және бұлардың мәнін нақты санмен көрсетуге талаптанушы ғалымдар да бар. Сөйтіп, қолданбалы тіл білімінің бүгінде «Статистикалық лингвистика» деп аталып жүрген жаңа саласының алғашқы нышандары мен белгілері міне осылардан басталады. Ал қазақ тіл біліміндегі бұл саланың бастауында профессор Құдайберген Қуанұлы Жұбанов тұрды. Мұны белгілі тілші-статист ғалымдар – Қ.Б.Бектаев, С.Мырзабеков т.б. қазақ тілінің статистикасы туралы зерттеулері мен мақалаларында атап көрсеткен. Мәселен, С.Мырзабеков «Қазақ тілін зерттеуде санды пайдалану» атты мақаласында: «Қазақ совет тіл білімінің негізін қалаушылардың бірі, совет дәуірінде қазақ тілін тұңғыш рет ғылыми түрде зерттеген маман-лингвист Қ.Жұбанов ретіне қарай сандық деректерді де пайдаланған», - десе [Мырзабеков 1975: 167], ғалым Қ.Б.Бектаев «Лингвистикалық статистикаға қатысты элементтерді – сандық деректер мен мәліметтерді, оны пайдалану нәтижесіндегі кейбір пікірлер мен тұжырымдарды біз профессор Қ.Жұбановтың еңбектерінен жиі кездестіреміз», - дейді [Бектаев, Белботаев, Смайылов 1990: 110]. Ғалымның сол кездің өзінде-ақ тіл білімінде сандық деректерді пайдаланудың қажеттілігіне ерекше ден қойғаны байқалады. Мәселен, ғалымның 1936 жылы жарық көрген «Қазақ тілінің грамматикасында» сөз мүшелерінің орны, дыбыстардың саны мен сапасы, түрлері туралы жазған мына тұжырымы көңілге қонымды: «Дыбыстың саны бар да, сапасы бар. Дыбыстың сапасы бірдей бола тұрып, саны әлденешеу бола береді. ...Бір тілде жұмсалатын дыбыстардың жалпы санының үшы-қиыры жоқ, бірақ оның бәрінің бірдей сапасы әр түрлі бола бермейді. Сапасы әр түрлі болатын дыбыстар да, көп болмайды». Сол сияқты «... Ілиястың небәрі 12 жол «Қойшы ойы» деген кішкене өлеңінде 488 дыбыс бар. Бірақ 488 дыбыс түрлі емес, мұнда небәрі 28-ақ түрлі дыбыс бар. Осы 28 түрлі дыбыстың бірі 10; бірі 50 рет қайталанып барып 488 болған. ... Қазақ тілінде де сапасы әр түрлі болатын дыбыстардың арнаулы саны бар» [Жұбанов 1999: 183].

Профессор Қ.Қ.Жұбановтың «Қазақ тілі жөніндегі зерттеулерінде» бұл жайлардан басқа да мәселелер, айталық: сөз тіркесінің формалары, сөйлем мүшелерінің орын тәртібі, буындар және ондағы дыбыс саны, буындарды оқыту әдістемесі т.б. мәселелер сөз болған тұста көптеген сандық (статистикалық) деректер мен мәліметтер келтірілген [Жұбанов 1999].

Шын мәніндегі статистикалық ізденістерді Қ.Жұбанов қазақ орфографиясын ғылыми негізде құру мақсатында қолданған. Ол әліпби құрамына енуге тиісті әріптердің санын қазақ тіліндегі белгілі бір фонеманы іс жүзіндегі қажеттігіне, оған деген ділгерлікке, қолданыс жиілігіне

бағындыра белгілеуді мақсат тұтқан [Бектаев, Белботаев, Смайылов 1990: 114].

Міне, бұл айтылғандардан проф. Қ.Қ.Жұбановтың тілді зерттеуде, оның заңдылықтарын қарауда есеп-санаққа, ретті жерінде математикалық ойлауға, оның тәсілдеріне сүйенгендігінің тағы да бір дәлелі.

Қ.Қ.Жұбанов еңбектеріндегі сандық деректерді пайдалану жайымен таныса отырып, мынадай қорытынды жасауға болады: проф. Құдайберген Қуанұлы Жұбановтың қазақ тіл білімі туралы зерттеулерінде тілді сандық деректер мен мәліметтер негізінде қараудың алғашқы нышандары бар. Бүгінде тіл білімінің бір саласына айналып отырған лингвистикалық статистиканы ғалым сол кездің өзінде-ақ тани білген. Тілдің сандық сипатынан байқалатын сапалық қасиеттерін оны зерттеуде пайдалануға болатын тиімді әдіс көздерінің бірі екенін де байқаған. Сөйтіп, ол сандық деректердің негізінде тілге қатысты бірсыпыра жайларды (дыбыстар, буындар, сөздер, сөйлем мүшелерінің орны т.б.) нақты анықтап, кейбір тың тұжырымдар жасаған. Бұлардың маңыздылығы сонда – олар қазіргі кезде бұл саладағы ізденуші-зерттеушілерге үлгі-тірек, бағыт-бағдар болып отыр [Бектаев, Белботаев, Смайылов 1990: 115-116].

Қазақ тілінің нормативті курсын жүйелеу жұмысына проф. Қ.Жұбановтан кейін белгілі тілші-ғалымдар: Н.Т.Сауранбаев, С.К.Кеңесбаев, М.Б.Балақаев, А.Ы.Ысқақов және т.б. ғалымдар атсалысты. Бастауыш және жоғары сыныптарға және педучилищелерге арналған қазақ тілі оқулықтарының авторлары (қалам иелері) тілдің буындық құрылымын, дыбыс физиологиясын, тілдің акценттік және екпіндік және т.б. мәселелеріне қатысты зерттеулерінде сан мен сапаны бөлмей, бірге қарастырады. Мәселен, С.К.Кеңесбаев қазақ тіліндегі буын типтерінің, М.Б.Балақаев қазақ тіліндегі синтаксистік сөз тіркестерінің табиғатын зерттеуде көптеген нақты мәтіндік материалдарды сандық қатынастар көмегімен қарастырған [Балақаев 1985; Кеңесбаев, Мұсабаев 1962]. Арнайы сандық талдауларды профессор Ғ.Мұсабаевтың зерттеулерінен де кездестіруге болады [Мұсабаев 1951]. Әрине, ғалымдардың мұндай ізденістері, яғни сандық қатынасқа көңіл бөлуі саналы түрдегі ықыластан гөрі тілдің ішкі табиғатына (сырына) қатысты туындаған деп пайымдауға болады.

Осы кезеңде қазақ тілінің грамматикалық форманттарын статистикалық жолмен зерттеуге айрықша көңіл бөліне бастаған болатын. Қазақ лексикографтары сандық деректі ұлы ақын Абай тілінің сөздігін түзу кезінде жүйелі түрде толық пайдаланды. Тұңғыш рет қазақ тіл білімінің тарихында бір ғана автордың (Абайдың) жазба түріндегі шығармалар тілі толық қамтылып, әр сөзі мен сөз тіркесінің қолдану жиілігіне көңіл бөлінді. Яғни Абайдың тіл байлығы, саны мен сапасы жайлы жайттар тілшілердің арнайы зерттеу нысанына айналды. Сөйтіп, орыстың ұлы ақыны Пушкин тілі сөздігі тәріздес – «Абай тілі сөздігі» жасалды [Абай тілі сөздігі 1968]. Бұл сөздікті әрі түсіндірме сөздік, әрі әліпби-жиілік сөздік деуге болады. Сөздердің қолдану жиілігін көрсету сөздік түзудің негізгі мақсаттарының бірі

болып саналды. Шығармадағы сөздердің жиілігін ескеру дәстүрлі картотека жасау жолымен іске асты.

1969 жылы Қазақстан Ғылым академиясы Тіл білімі институтында «Статистическое и информационное изучение тюркских языков» атты Бүкілодақтық ғылыми жиынның өтуі қазақ тіл білімі үшін елеулі құбылыс болды. Осы ғылыми форумның нәтижесі ретінде 1970 жылы Тіл білімі институтында статолингвистика саласының белгілі ғалымы Қалдыбай Бектаевтың жетекшілік етуімен «Тіл статистикасы және автоматтандыру» атты ғылыми топ құрылды. Осы кезеңнен бастап қазақ тілін статистикалық зерттеу ісі жүйелі сипатқа ауысты. Бірнеше жас ғалым статистикалық лингвистика саласынан кандидаттық диссертация қорғап, олардың ғылыми мақалалар жинағы, монографиялары, оқу құралдары жарық көрді. Кеңес Одағы кеңістігіндегі ғылыми конференцияларға кеңінен жол ашылды. Қазақстан статолингвистер зерттеулерінің нәтижесі ретінде 1973 жылы «Қазақ тексінің статистикасы» атты ғылыми жинақ жарық көрді [Қазақ тексінің статистикасы 1973]. Бұл жинақ сол жылға дейінгі қорғалған (және дайындалған) кандидаттық диссертациялардың негізгі мазмұнын ашады деуге болады.

1973-1974 жж. Р.Г.Пиотровский мен Қ.Б.Бектаевтың «Математические методы в языкознании» атты екі бөлімнен тұратын, жоғары оқу орындарына арналған оқу құралының Қазақ мемлекеттік университетінен жарық көруі де осы жаңа саланың жетістігі еді [Бектаев, Пиотровский 1973; Бектаев, Пиотровский 1974].

Тілді зерттеу тәжірибесінде әр тілді жалпы типологиялық тұрғыдан бағалауда қолданылатын статистикалық критерийлердің үлес салмағын «өлшеуге» (анықтауға) қажетті жалпылама сандық дәреже әлі де ғалымдарды толық қанағаттандырмай келеді. Міне, осы бос кеңістікті толтыруға арналған іргелі ғылыми зерттеу ретінде 1978 жылы басылып шыққан Қ.Б. Бектаевтың «Статистико-информационная типология тюркского текста» атты монографиясын атауға болады [Бектаев 1978]. Бұл зерттеу түркі мәтіндері үшін ықтималдық және ақпараттық үлгілеу әдістерін жүйелі түрде қолдануға арналды. Ғалым түркі тілдерінің типологиялық ерекшеліктерін үнді-еуропа тілдерімен салыстыра зерттеуді мақсат етіп қойып, оны формалды статистикалық-ықтималдық әдіс арқылы іске асырды.

Осы кезеңдік зерттеулерге тән мәселелер – ол тілдің фонетика мен фонология саласы, сөз бен сөзтұлғаның буындық құрылымы, қазақ мәтінінің түрлі функционалды стилінің лексика-морфологиялық құрылымы, қазақ тілінің ақпараттық сипаты және т.б. кешенді проблемалар болатын. Негізгі зерттеуге қажетті статистикалық құрал ретінде электронды-есептеу машиналары (ЭЕМ) көмегімен жасалатын әліпби-жиілік сөздік, жиілік сөздік және кері әліпби жиілік сөздіктер материалдары болды. Қазақ тілін статистикалық әдіспен зерттеу және оны автоматтандыру мәселелері де Одақтық дәрежеде қарастырылды. Отандық түркітану ғылымы үшін ортағасырлық жазба ескерткіші «Кодекс Куманикус» мәтіні тұңғыш рет ЭЕМ жадына енгізіліп, ол бойынша автоматты түрде «Куманша-қазақша жиілік

сөздігінің» алынуы қазақ тілі тарихы үшін елеулі оқиға деуге болады [Құрышжанов, Жұбанов, Белботаев 1978].

Жоғарыда сөз болған қазақ тілі мәтіндерін статистика тәсілімен зерттеуге арналған кандидаттық диссертациялар тақырыбы 1979 жылы «Ғылым» баспасынан жарық көрген М.О.Әуезовтің «Абай жолы» романының жиілік сөздіктеріне тікелей не салыстырмалы тұрғыда қатысты болды [Бектаев, Жұбанов, Мырзабеков, Белботаев 1979]. Ал ЭЕМ көмегімен алынған және 1995 жылы басылып шыққан «М.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері» ұлы жазушының 100 жылдық мерейтойына арналған Қазақстандық статолингвистерінің арнайы сыйы деуге болады [Бектаев, Жұбанов, Мырзабеков, Белботаев 1995]. Жазушы тілінің лексикалық байлығын, автордың сөз қолданудағы стильдік тәсілінің өзіндік ерекшелігін, қазақ тілінің көркемдік және бейнелеуші мүмкіндіктерін меңгеру дәрежесін осы аталған лексикографиялық еңбек бойынша бағалауға болады. Осымен қатар бұл сөздік жалпы қазақ лексикасын статистикалық құрылым тұрғысынан зерттеуге бай тілдік материал бола алатыны сөзсіз.

А.Ахабаевтың қазақ тілін нормативті тұрғыда тануға, қазіргі қазақ публицистикасы тілінің лексика-морфологиялық құрылымын статистикалық талдауға арналған ғылыми жұмысы да маңызды зерттеу болып табылмақ [Ахабаев 1971]. Статистикалық талдау негізінде А.Ахабаев сөздердің жиілік сипаты негізгі объективті критерий болатындығын сөз етеді. Жалпы алғанда, бұл ғылыми жұмыс 1965-1966 жж. газет мәтіндері мен «Абай жолы» романы (2-ші кітап) мәтініндегі зат есім мен есімдік жалғауларының форма-варианттарын статистика-морфологиялық тұрғыда салыстыра талдауға арналып, соның негізінде қазақ тілінің грамматикалық тұлғаларының нұсқалары стильдік айырым белгілер болатыны анықталған.

Статистика-морфологиялық зерттеудің келесі бір нысаны – қазіргі қазақ тілі етістік сөздерінің туынды тұлғаларын құрайтын негіз етістік пен оларға жалғанатын аффикстер. Осы зерттеуді М.Әуезовтің «Абай жолы» романындағы туынды түбір етістіктерінің құрылымына статистикалық талдау жүргізуді ғалым С.Мырзабеков жүргізді. Ғалымның сандық деректері бойынша романда кездесетін барлық түбір етістіктің 80 пайызы – туынды етістіктер екені және етістіктің туынды түбірлері негізгілерден төрт есе көп қолданылатыны анықталды [Мырзабеков 1973]. Етістік құрылымына жасалған лингва-статистикалық талдаудың нәтижелері қазіргі қазақ тіліндегі етістік сөздерді танып-білуге мүмкіндік беретіні сөзсіз. Қазақ тіліндегі жалғаулар мен олардың құрылымдық тұлғалары, терминдер мен кірме сөздер, сөз тіркестері статистикасы сияқты мәселелер А.Белботаевтың ғылыми ізденістерінде қарастырылды [Белботаев 1992].

Тіл ерекшелігі мәтіндердің жанрлық (стильдік) түрлеріне қарай да айырым табатыны белгілі. Осындай айырымдарды қазақ тілінің әр түрлі жанрларына қатысты статистикалық жолмен тағайындауды қазақстандық статистика тобының ғалымдары 70-80-жылдары өз мақсаты етіп қойды. Мысалы, Қ.Молдабеков пен Б.Е.Қалыбековтер балалар әдебиеті мен бастауыш сыныптарының оқулықтарын зерттеуді өз міндеттеріне алса,

А.Р.Зекенова – М.Әуезовтің драмалық шығармалар тілін, А.Ахабаев – газет тілін, А.Белботаев – ғылыми-техникалық стиль, математикалық стиль тілдерін өз зерттеу нысандары етіп алып, алғашқы ғылыми нәтижелерге қол жеткізді [Ахабаев 1971; Белботаев 1992; Белботаев 1973; Молдабеков 1985; Зекенова 1976; Қалыбеков 2003].

Қазақ тілінің фонетика саласы да статистика әдісінің назарынан тыс қалған жоқ. Фонетиканың дәстүрлі (есту) жәие құралды-тәжірибелік зерттеу әдісі бойынша фонологиялық және графикалық жүйелерге қатысты мәселелерін ықтималды-статистикалық әдіс-тәсілдермен шешу тиімді деп танылды [Джубанов 1979; Джубанов, Джунисбеков 1980; Татубаев 1969; Кенесбаева 1969; Бектаев 1973].

Қазақ тіл білімінің қолданбалы саласының тың проблемалық мәселесі – қазақ мәтінінің әр әріпке (дыбысқа) шаққандағы ақпараттық өлшемін, яғни *энтропияны* анықтау. Осындай зерттеулер белгілі ғалым Қ.Б.Бектаев пен оның шәкірті Д.А.Байтанаеваның қазақ мәтініне қатысты ақпараттық-статистикалық ғылыми ізденістері арқылы іске асырыла бастады. Табиғи тілдің белгілер жүйесі үш түрлі деңгейде – *синтактикалық, семантикалық* және *прагматикалық* аспектілерде қарастырылатыны мәлім. Мұндағы синтактика деңгейі әр түрлі белгілер жүйесінің синтаксисін, яғни сөз мағынасынан тыс, белгілердің тіркесімдік құрылымын және олардың туындау ережелерімен қатар белгілер құрылымының ішкі қасиеттерін қарастырады. Ал белгілер жүйесінің сыртқы құрылымдық қасиеттерімен семантика мен прагматика айналысады. Дәлірек айтсақ, егер *семантика* белгілер жүйесін мағыналық көзқарас, яғни мәтін мазмұны тұрғысынан танып-білумен шұғылданса, *прагматика* белгілер жүйесінің қабылдаушыға қатынасын анықтау мәселелерімен айналысады.

Қазіргі кезде ақпарат аясын теориялық және технологиялық тұрғыда зерттеудің синтактикалық деңгейі жақсы дамып келеді. Бұған К.Шеннонның символдар жүйелілігі мен комбинаторикасын қарастыратын статистикалық байланыс теориясы негіз болады. Бұл теория бойынша табиғи тілде жазылған мәтін белгілі бір код құрайтын дискретті белгілердің тізбегі ретінде қарастырылады. Ал мұндай мәтінді қабылдау кезінде дискретті белгілердің қай-қайсысының да пайда болуы белгісіздіктің шамасымен, яғни энтропиямен өлшенеді. Қ.Б.Бектаев пен Д.А.Байтанаеваның ғылыми ізденістерінде қазақ мәтінінің ең кіші бірлігі – әріпке түсетін энтропияның орта шамасын және қазақ тілінің толық және стильдік (жанрлық) айырымдарын анықтайтын жазба мәтіндеріндегі энтропия мен артық ақпарат шамаларын анықтау негізгі мақсат ретінде қойылды. Осымен қатар қазақ тілі бойынша анықталған осындай сипаттамаларды синтактикалық деңгейде үнді-еуропа тілдері бойынша алынған белгілі деректермен салыстыра зерттеулер де іске асты. Осы аталғандармен бірге Д.А.Байтанаева қазақ жазба мәтінінің буын бірлігінің статистика-ақпараттық қасиеттерін төжірибелік мәліметтер негізінде айқындайтын статистикалық зерттеулер де жүргізді [Байтанаева 1969]. Кезінде Бүкілодақтық статистика тобы мүшелігінде Қазақстаннан басқа да түркі

тілдес мемлекеттердің статист-тілшілері болды. Осының нәтижесінде 1988 жылы К.Дыйқановтың қырғыздың «Манас» эпосының 2 бөлімнен тұратын жиілік сөздігі жарық көрді. Ж.Жетешиков қырғыз публицистика тіліндегі зат есімдердің сөз өзгертуші аффикстерінің статистикалық деректерін қазақ тілінің осындай мәліметтерімен салыстыра зерттеді [Жетешиков 1980]. С.И.Ибрагимов ғылыми-техникалық стильдер мәтіндері бойынша қырғыз тілінің зат есімдерінің синтаксисті-функционалды кластарын бөлек қарастырып, олардың басқа сөз таптарымен синтаксистік байланыс түрлерін статистикалық жолмен қарастырды [Ибрагимов1969].

Қырғыз тілі бойынша жүргізілген іргелі ғылыми-зерттеу жұмыстары қатарына Т.Садықовтың «Проблемы моделирования тюркской морфологии» (аспект прождения киргизской именной словоформы) атты монографиясын [Садықов 1987] қосуға болады. Бұл жұмыста қырғыз тілінің алдын ала берілген лексика-грамматикалық ақпараты бойынша есім сөзтұлғалардың туындау үлгісін (моделін) құру көзделді. Ол үшін қырғыз тілінің морфологиялық құрылымына жүйелі түрде талдау жасалды [Садықов 1987].

Қазіргі өзбек тілінің статистикалық құрылымын зерттеушілердің бірі С.А.Ризаев болды. Ол өзбек әдеби тіліндегі екіфонемдік тіркесім жағдайын қарастырып, олардың қатынастық жиілікке әсер ететін факторларын анықтады. Публицистика, ғылыми және кәркем әдебиет (балалар ертегісі) стильдерінің түрлі жанрларындағы фонеманың сандық қатынастарын қазақ тіліндегі осындай деректермен статистика жолымен салыстыра зерттеп, олардың өзіне тән ерекшеліктері негізінде фонемалық топтарын бөліп алуға мүмкіндік туды. Қазақ тілі мен өзбек тілдерінің фонемалық спектрлерін статистикалық талдауға А.Жүнісбеков пен Д.Маматов өз зерттеулерін арнады [Джунисбеков 1972; Маматов 1982]. Бұл жұмыстарда қазақ тілінің дауысты, ал өзбек тіліндегі күрделі дауыссыз дыбыстардың спектрлері қарастырылды.

Осының негізінде дербес фонемалар мен олардың позициялық варианттары жайлы статистикалық деректер алынды. Осы тәріздес зерттеу жұмыстары қазақ, қарақалпақ пен әзірбайжан тілдері бойынша да жүргізілді [Насыров, Айымбетов 1980; Вейсалов, Керимов, Сафаров 1981; Байтанаева 1985]. С.А.Ризаев өзбек тілінің буын типтерін және олардың сөздердегі позициялық орналасуы мен ашық, жабық буындардың қатынастық сипаттары негізінде сөздердің тұрақты (канондық) тұрпатын анықтады. Мұндай зерттеулер балалар әдебиеті бойынша да жүргізілді [Ризаев 1972, 1980].

Қазіргі татар әдеби тіліндегі буын саны мен олардың буындық құрылым типтері, комбинаторлық қасиеті, сөздердегі буындардың қолдану жиілігі жайлы мәселелерді Т.И.Ибрагимовтың ғылыми ізденістерінен кездестіруге болады [Ибрагимов 1966, 1970]. Мәселен, татар тілі сөздеріндегі алдыңғы буын өзінен кейін келетін буын жайлы ақпараттың 47 пайызын өз бойында ұстайтындығы анықталды. Мұндай деректер тілдің құрылымдық және фонетикалық құрылымы жайлы көптеген мәлімет беретіні белгілі. Қ.Б.Бектаев көркем әдебиет, публицистика, ғылыми-техникалық мәтіндер

мен 2 томдық «Қазақ тілінің түсіндірме сөздігі» (1959, 1961 жж.) материалдары бойынша ЭЕМ арқылы қазақ тілі буындарының жиілік сөздігін жасады [Бектаев 1973]. Ал Д.А.Байтанаева қазақ мәтіндеріндегі фонемалардың ғана емес, сөздердегі буындардың да статистика-ақпараттық қасиеттерін тәжірибе жүзінде қарастырды [Байтанаева 1969].

Статистикалық әдіс түрлі тілдердегі грамматикалық тұлғаларды салыстыра зерттеуде де елеулі нәтижелерге ие болды. Мәселен, Т.М.Гарипов башқұрт, қазақ, татар, чуваш түркі тілдерін лексика-семантикалық тұрғыда талдап, олардың сандық және сапалық ерекшеліктерін анықтауды мақсат етті [Гарипов 1972]. Сол сияқты туыстас емес тілдер қазақ және орыс, қазақ және ағылшын тілдерін салыстырғандағы грамматикалық формалар да статистикалық зерттеу аясынан тыс қалмады [Ишанов, Садиков, Мырзабеков 1969; Ахметов 1969; Ергалауов 1969].

1987 жылы Қаз. ССР «Наука» баспасынан шыққан (осы жолдар авторының) «Квантитативная структура казахского текста» және 2002 ж. жарық көрген «Основные принципы формализации содержания казахского текста» атты монографияларында қазақ тілінің заманға сай өзекті зерттеу мәселелері сөз етілді. Бірінші аталған ғылыми еңбекте қазақ тіліндегі мәтіндерді автоматты түрде электронды-есептеу машиналары (ЭЕМ) арқылы статистикалық әдіспен зерттеу мүмкіндігі мен сөз таптарының ықтималды-үлестірімді заңдылықтары зерттелсе, екінші монографияда – қазақ мәтінінің мәтін лингвистикасы тұрғысынан мазмұнын формалдау жолдары және мәтіннің туындау (порождение) проблемалық мәселелері қарастырылады [Джубанов 1987; Жубанов 2002].

Сонымен, жоғарыда сөз болған ізденістер жеке автор шығармалары, функционалдық стильдер не толық әдеби тіл қамтылған мәтіндер бойынша статистикалық әдіс-тәсілдермен зерттеу тәжірибесінде әр түрлі жиілік сөздіктер жасаудың өзекті мәселе (актуалды) екендігін көрсетті. Себебі олардың негізінде аса маңызды және күрделі мәселелерді: тілді лексика-грамматикалық нормалау, түркі тілдерін өзара салыстырмалы-типологиялық зерттеу және олардың компьютерлік қорын жасау мәселелерін шешуге зор мүмкіндік жасалады.

Статистикалық лингвистика саласына қатысты зерттеулер жайлы жоғарыда келтірілген қысқаша шолудан ең алдымен байқайтынымыз, мұндағы зерттеу аясының (нысанының) кеңдігі және бұл әдістің әмбебаптық сипаты. Солай бола тұра, біз қарастырған зерттеу жұмыстарының нәтижелері мен сапасы, деңгейі бірдей дәрежеде емес. Егер қазақ, қырғыз, әзірбайжан және өзбек тілдеріндегі ізденістер диссертациялық деңгейдегі іргелі зерттеулер қатарына жатса, ал кейбір басқа тілдерге қатысты статистикалық зерттеулер жаңадан ғана бастама алуда және олар жайлы тезис түріндегі қысқаша хабарламалар бойынша аз ғана пікір айтуға болады. Сол сияқты мұнда тіл білімінің барлық салалары да біркелкі қамтылмаған. Мәселен, қай түркі тілін алсаңыз да, олардың құрылымын статистикалық лингвистика тұрғысынан зерттеуде синтаксис пен семантика салалары тиісті дәрежеде қамтылмаған деуге болады. Әрине, аталған әдіс

түркі тілдері үшін «жаңа» болғанымен, кейбір түркі тілдері үшін бұл әдіс тұрақтанып, түркітанушы ғалымдардың зерттеулерінен елеулі орын алған. Басқа түркі тілдерін айтпағанда, қазақ тілінің көптеген кезек күттірмейтін мәселелері, мәселен лексикография саласы бойынша компьютерлік қор (база) жасау, қазақ тілінен басқа тілге, не керісінше компьютерлік аударма жасау, ақпарат мазмұнынан қысқаша реферат, аннотация алу және қолданбалы лингвистика саласының басқа да мәселелері өз шешімін күтуде.

Әдебиеттер:

1. Абай тілі сөздігі. - Алматы, 1968. 734 с.
2. *Ахабаев А.* Статистический анализ лексико-морфологической структуры языка казахской публицистики. Автореф. дис. ... канд. филол. наук. -Алма-Ата, 1971. - С. 24.
3. *Ахметов Е.Г.* Статистическая характеристика родительного падежа имени существительного в английском и казахском языках // Тезисы докладов и сообщений всесоюзного семинара «Статистическое и информационное изучение тюркских языков». Алма-Ата, 1969. -С. 51-54.
4. *Байтанаева Д. А.* Информационные характеристики казахского текста. - Алма-Ата, 1985.- С. 18.
5. *Байтанаева Д.А.* Энтропия распределения слогов казахского письменного текста // Тезисы докладов и сообщений всесоюзного семинара «Статистическое и информационное изучение тюркских языков». - Алма-Ата, 1969. С. 80-82.
6. *Байтанаева Д.А.. Бектаев К.Б.* Энтропия казахского текста // Статистика казахского текста. -Алма-Ата, 1973. -С. 664-696.
7. *Балакаев М.* Современный казахский язык. -Алма-Ата, 1985.
8. *Бектаев Қ., Белботаев А., Смайылов Т.* Қазақ тіл біліміндегі лингвистиканың алғашқы көріністері // Құдайберген Жұбанов және қазақ совет тіл білімі. Алматы «Ғылым», 1990. - Б. 108-115.
9. *Бектаев К.Б.* Алфавитно-частотный словарь слогов казахского языка // Статистика казахского текста. -Алма-Ата, 1973. - С. 566-612.
10. *Бектаев К. Б.* Статистико-информационная типология тюркского текста. – Алма-Ата. 1978. -183 с.
11. *Бектаев Қ.Б., Жұбанов А.Қ., Мырзабеков* С., *Белботаев А.Б.* М.О.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері. - Алматы-Түркістан, 1995. -346.
12. *Бектаев Қ.Б.. Жұбанов А.Қ., Мырзабеков С, Белботаев А.Б.* М.Әуезовтің «Абай жолы»романының жиілік сөздігі. -Алматы, 1979. -334.
13. *Бектаев К.Б., Пиотровский Р.Г.* Математические методы в языкознании. Теория вероятностей и моделирование нормы языка. - Алма-Ата. 1973. - 281 с.
14. *Бектаев К.Б., Пиотровский Р.Г.* Математические методы в языкознании. Математическая статистика и моделирование текста. - Алма-Ата. 1974. - 334 с.

15. *Белботаев А.Б.* Ғылыми-техникалық стильдегі сын есім аффикстерінің қолданылуы туралы; Математика тексінің алфавитті-жиілік сөздігі // Қазақ текенің статистикасы. - Алматы: Ғылым, 1973. - Б. 225-230. 543-552.

16. *Белботаев А. Б.* Лингвостатистические характеристики частей речи казахского текста. Автореф. дис. ... канд. филол. наук. Алма-Ата. 1992. -С. 32.

17. *Вейсалов Ф. Е., Керимов Я.К.. Сафаров М.Г.* Комбинаторика фонем современного азербайджанского языка // Вопросы азербайджанского языкознания. - Баку, 1981.

18. *Виноградов В.В.* Современный русский язык. Введение в грамматическое учение о слове. -1938. -Вып. 1.-160с.

19. *Гарипов Т. М.* Лексикостатистические отношения тюркских языков Урало-Поволжья //Языковые контакты в Башкирии. - Уфа, 1972. - С. 248-267.

20. *Головин Б.Н.* Язык и статистика. - М., 1971. - 191 с.

21. *Джубанов А.Х.* К вопросу о графемной статистике казахского текста // Вопросы казахской фонетики и фонологии. -Алма-Ата. 1979. - С. 79-86.

22. *Джубанов А.Х.* Квантитативная структура казахского текста (опыт лингвистического анализа на ЭВМ). -Алма-Ата, 1987. 147 с.

23. *ДЖубанов А.Х.. Джунисбеков А.* Преобразование с помощью ЭВМ казахских текстов в фонетическую запись // Материалы семинара «Статистическая оптимизация преподавания языков и инженерная лингвистика». - Чимкент. 1980. -С. 317-319.

24. *Джунусбеков А.* Гласные казахского языка. -Алма-Ата. 1972. -С. 94.

25. Ергалауов А. Статистическая характеристика форм страдательного залога в английском и казахском языках // Тезисы докладов и сообщении всесоюзного семинара «Статистическое и информационное изучение тюркских языков». -Алма-Ата, 1969. - С. 62-63.

26. *Жетешиков Ж.* Статистический анализ словоизменительных аффиксов имен существительных киргизского языка (на материале газетных текстов) // Материалы семинара «Статистическая оптимизация преподавания языков и инженерная лингвистика». - Чимкент, 1980. - С. 25.

27. *Жубанов А.К.* Основные принципы формализации содержания казахского текста. -Алматы. 2002. - 250 с.

28. *Жубанов Қ.* Қазақ тілі жөніндегі зерттеулер. -Алматы: Ғылым. 1999. -581 б.

29. *Зекенова А. М.О.*Әуезов пьесаларындағы есім сөздердің морфологиялық құрылысы жайында // Известия АН КазССР. Сер. филолог. — 1976. - С. 56-62.

30. *Ибрагимов С.И.* Некоторые статистические характеристики имен существительных киргизского языка // Тезисы докладов и сообщений всесоюзного семинара «Статистическое и информационное изучение тюркских языков». - Алма-Ата, 1969.-С. 34-38.

31. *Ибрагимов Т.Н.* Изучение слогов и структуры их сочетаний в

татарском литературном языке. - Казань, 1970.

32. *Ибрагимов Т.И.* Некоторые статистические данные о слогах татарского языка: вероятностные методы и кибернетика // Учен. зап. Казанского ун-та. - Вып. 4. -Т. 25. Кн. 6. -1966. -С. 74-78.

33. *Ишанов К.И.. Садиков А. В., Мырзабеков С.* Частотно-сопоставительная характеристика падежных форм русского и казахского языков // Тезисы докладов и сообщений всесоюзного семинара «Статистическое и информационное изучение тюркских языков». -Алма-Ата, 1969. -С. 56-58.

34. Қазақ тексінің статистикасы. -Алматы, 1973. - 731 б.

35. *Қалыбеков Б.Е.* 50-жылдардағы бастауыш сынып оқулықтарының лексикалық жүйесі мен морфологиялық құрылымының статистикасы. Афтореф. дис. ...канд. наук. - Алматы, 2003. - 25 с.

36. *Кенесбаева С.С.* К вопросу исследования протетического звука в словах арабского происхождения статистическим методом // Тезисы докладов и сообщений всесоюзного семинара «Статистическое и информационное изучение тюркских языков». -Алма-Ата. 1969. - С.43-44.

37. *Құрышжанов А.Қ., Жұбанов А.Қ., Белботаев А.Б.* Куманша-қазақша жиілік сөздік. -Алматы: Ғылым, 1978. -277 с.

38. *Маматов Д.* Количественный анализ сложных согласных // Опыт экспериментального и структурного изучения языка. - Ташкент, 1982. - С. 108-111.

39. *Молдабеков К.* Лингвостатистические исследования казахских текстов для младших школьников (на материале текстов учебников начальных классов и литературы для детей). Дисс. канд. филол. наук. -Алма-Ата, 1985. -259с.

40. *Мұсабаев Ғ.* Қазіргі қазақ тіліндегі сын есімнің шырайлары. - Алматы, 1951. - 90 б.

41. *Мырзабеков С.* Қазақ тілін зерттеуде санды пайдалану // Қазақ тілі грамматикасы бойынша зерттеулер. -Алматы, 1975.

42. *Мырзабеков С.* Статистико-лингвистический анализ структуры глагола современного казахского языка. Автореф. дис. ... канд. филол. наук. - Алма-Ата, 1973.-С. 32.

43. *Насыров Д.Х., Аймбетов А.* О статистике фонем каракалпакского языка // Материалы семинара «Статистическая оптимизация преподавания языков и инженерная лингвистика». - Чимкент, 1980. -С. 273-274.

44. Пиотровская А.А. Автоматическое приведение именных словоупотреблений к канонической форме //НТИ. Сер. 2. 1977. №1. - С.32-36.

45. *Пиотровский Р.Г.* «Экстралингвистические и внутриязыковые вопросы при переработке текста в системе «человек-машина-человек» // Вопросы социальной лингвистики. -Л.. 1969. -С. 40-64.

46. *Ризаев С.А.* Из опыта исследования статистической структуры слога в узбекском языке // Общественные науки в Узбекистане. - Ташкент, 1972. -№1.

47. *Ризаев С.А.* Статлингвистический анализ структуры слова в современном узбекском языке // Материалы семинара «Статистическая оптимизация преподавания языков и инженерная лингвистика». -Алма-Ата.

1980. -С. 180-182.
48. *Садыков Т.* Проблемы моделирования тюркской морфологии (аспект порождения киргизской именной словоформы). - Фрунзе, 1987. -С. 121.
49. *Струве П.Б.* Кто первый указал на применение статистики к филологическим исследованиям? // Изв. Российской АН. Сер. 6. - 1918. -№13.
50. *Татубаев С.С.* Статистические подходы к исследованию певческой фонетики казахского языка // Тезисы докладов и сообщений всесоюзного семинара «Статистическое и информационное изучение тюркских языков». - Алма-Ата. 1969. - С. 38-40.

ҚАЗАҚ МӘТІНІН ТАЛДАУ МЕН ЖИНАҚТАУДЫҢ АВТОМАТТЫ ЖҮЙЕЛЕРІН ҚҰРУ МӘСЕЛЕСІ

«Академик Ш.Ш. Сарыбаев және тіл білімінің мәселелері» атты халықаралық ғылыми-теориялық конференция материалдары. Алматы. Дайк-Пресс. 2005 ж., 693-699 б.б.

Қазіргі қоғамдағы ақпарат ағымының үнемі ұлғаю сипаты тұтынушы мүмкіндігімен сай келе бермейді. Мәселен, Кеңес Одағы кезіндегі статистикалық мәліметтерге қарағанда, халық шаруашылығы бойынша алынған ауыспалы ақпарат көлемінің әрқайсысы 500 беттен тұратын 25 миллион кітап санымен есептелсе, жазба құжаттардың жылдық саны – 60 миллиардпен және әлемдік шеңберде, тек патенттік құжаттардың жылдық өсу қарқыны ғана 0,5 миллион мен 1 миллион аралығында болғаны белгілі. Басқаша айтқанда, бір маман ғылыми-техникалық ақпаратпен ғана хабардар болуы үшін 1 минутта 2 мың баспа табақ материалмен таныс болуы қажет екен. Сол сияқты әрбір педагог, инженер, ғалым өз саласы бойынша әлемдік әдебиеттерді күн сайын бақылап отыруы үшін ғана кемінде 1500 беттен тұратын мәтінді оқып шығуы қажет болады [1].

Әрине, мұндай көлемдегі ақпараттарды игеру адам мүмкіндігімен сай емес екені белгілі. Сондықтан Кеңес Одағы шеңберінде барлық ашылған жаңалықтар мен ұсыныстардың тек төрттен бірі ғана соны жаңалыққа саналып, ал басқалары, негізінен, қайталаулар болып табылған [2]. Мұндай кейбір құнды ақпараттардың мамандар назарынан тыс қалуы мемлекет деңгейіндегі экономикалық шығынға әкелетіні сөзсіз.

Аталған қиыншылықтармен күресу жолдарының алғашқылары ретінде елімізде ашылып жатқан кітапханалар мен ақпараттық зерттеу институттарын (ВИНИТИ, ЦНИИПИ және т.б.) атауға болар еді. Сол сияқты шет тілінен аударма, реферат пен аннотация жасау істері және қажетті мәліметтерді дер кезінде тауып алу жұмыстары дәстүрлі әдістермен (автоматсыз жолмен) орындалуда. Бірақ ақпараттық қызметтің мұндай түрі зерттеуші-ғалым уақытын үнемді пайдалануға кері әсер ететіні дәлелденіп отыр. Мысалы, статистердің деректеріне қарағанда, дәстүрлі әдіс бойынша

ақпаратты тұтынушы адамның пайдалы уақытының 80 пайызы қажетті мәліметтерді іздеу, талдау және сұрыптау үшін ғана кетеді екен.

Міне, осы себептерден ақпараттық қызметті жаңаша жолға қою үшін тілдік мәліметтерді «өңдеудің» (обработка информации) жаңа әдістерін іздестіру қажет болып отыр. Яғни табиғи тілдегі ақпарат ағымын «тізгіндеудің» ең бір ұтымды жолы ретінде қазіргі заманда жаңа технологияны (электронды-есептеу техникасын) – компьютер мүмкіндіктерін молынан пайдалануды айтуға болады.

Қазіргі кездегі компьютерлер мынадай үш түрлі ақпарат жүйесі үшін ең негізгі техникалық база болып саналады:

1. Автоматтандырылған аударма жүйесі (ААЖ) мен аннотациялау және рефераттау жүйелері (АРЖ).

2. Автоматтандырылған аударма жүйесі (ААЖ) мен ақпараттық-іздеу (АІЖ) жүйесі.

3. Автоматтандырылған сараптама жүйесі (АСЖ).

Осы аталған жүйелердің ұтымды жері сол табиғи тілдің функционалды моделінің құрылымдық сапасына тікелей байланысты болуында [3].

Жоғарыда аталған үш түрлі жүйеде мәтінге қатысты орындалатын амалдар, яғни аударма жасау, рефераттау мен аннотациялау, ақпаратты іздеу мен мәтінді туындату (порождение текста) проблемасы әр сөздің мағынасы арқылы сөйлемнің мән-мағынасын (смысл) ашуға, жинақталған белгілі нәтижелер негізінде мәтін мазмұнын айқындау принципіне негізделеді. Бірақ сөйлемнің мағынасы (смысл) оның құрамындағы сөз мағыналарының, ал мәтін мазмұны – сөйлемдер мән-мағыналардың ғана жиыны емес екендігі айқын. Сондықтан аталған жүйелерді құруда (модельдеуде) адам қызметінің мәтінді туындатуына қатысты шығармашылық жағын бағыт-бағдар ету қажет.

Тіл қызметінің формальды моделін (үлгісін) құру синтетикалық процесс екендігін ескерсек, аталған модельді құруда лингвистермен бірге психолингвистер, психологтар, логиктер, философтар, кибернетиктер, математиктер және т.б. өзара бірлесе қызмет етуі қажет.

Зерттеуші ғалымдар аталған формальды модельді құру барысында мынадай негізгі үш аспектіні ұстаным еткені жөн:

1) тілді зерттеу мен сипаттауда адресант (айтушы) пен адресат (тыңдаушы, қабылдаушы) факторы есепке алынбайды;

2) тілді зерттеуде адресат (тыңдаушы, қабылдаушы) факторы ғана ескеріледі;

3) тілді зерттеуде оның құрылымдық жағы, яғни адресант (айтушы) факторы ғана ескеріледі.

Таза теориялық зерттеу барысында аталған аспектілер бірдей дәрежеде орын алады да, ал «мәтінді өңдеуге» (обработка текста) қатысты іске қосылатын автоматты жүйелерде екінші және үшінші реттегі аспектілер негізгі бағыт болып саналады.

Негізінен алғанда, осы бағыттағы зерттеулер көбінесе, қиындасқан мәтіндерді (связные тексты) талдау (анализдеу) проблемасына арналып, ал

мәтінді туындату (порождение текста) модельдерін құру жинақтау (синтездеу) әдісімен іске асып жүр [4]. Бірақ, қазақ тіліне қатысты қиындасқан мәтіндерді жинақтау (синтездеу) мәселесі әлі де болса арнайы зерттеулерді қажет етуде. Мұндай зерттеулер дәстүрлі әдістер арқылы байқала бермейтін жаңа заңдылықтарды ашуға мүмкін жасайтыны сөзсіз. Практикалық тұрғыдан алғанда, аталған зерттеулердің нәтижесі мәтіндердің динамикалық заңдылықтарын анықтауда және қазақ мәтінін туындату мәселесіне қатысты аса маңызды деуге болады. Түрлі автоматтандырылған басқару жүйелерінің даму барысы адам мен ЭЕМ арасындағы коммуникацияны табиғи тіл негізінде жүргізуді талап етеді. Ал ондай автоматтандырылған жүйелердің «адамша сөйлеуі» оның «адамша ойлайтын» қабілетінің болуын қажет етеді [5]. Мұндай «ойлау» механизмі іске асса, мәтін бойындағы неше түрлі тұрақтанған (бір үлгідегі) сипаттауларды талдау және жинақтау жұмыстары автоматты түрде орындалар еді. Бұл жердегі «тұрақтанған сипаттауларға» (стандартное описание) мәтін ішіндегі арнайы түрдегі баяндау сипатын, азаматтық не әскери құжат мәтіндерінің ерекшелік сипатын және т.б. жатқызуға болады.

Мәтінді туындату заңдылықтарын меңгерудің практикалық маңызды жақтарының бірі – орта және жоғары білім беру жүйесін жаңаша ұйымдастыру мәселесі. Аталған мәселелерді шешу үшін әр түрлі мәтіндердің құрамдық сипатын, олардың туындау заңдылықтарын біліп, мәтін ішіндегі негізгі және көмекші тілдік элементтерді айыра білу керек. Ал бұл, бір жағынан, оқушыларға қажетті деген оқу материалын сапалы меңгеруіне, екіншіден, мамандардың жаңа оқулықтар мен оқу құралдарын құрастыруына игі ықпал етеді.

Жоғарыда аталған мәселелер барлық түркі тілдерін, соның ішінде қазақ тілін зерттеу ісіне де тікелей қатысы бар. Әсіресе қазақ тілінің мемлекеттік тіл дәрежесін алуы, оның халық шаруашылығын басқаруда, оқыту мен тәрбие жұмыстарында, қазақ халқының мәдени өмірі мен егемен елімізді қазақ тіліндегі ақпарат ағымымен қамтамасыз ету мәселесінде маңызды болмақ.

Ұлттық ғылым салаларының қарқындап дамуы ең алдымен халық шаруашылығын қазақ тілінде басқаратын автоматтандырылған жүйе құруды талап етеді. Жаңа технологияның негізінде сапалы түрде жұмыс атқаратын мұндай тілдік жүйе құру мәселесі өмір талабына сай және мемлекеттік дәрежеде іске асуы қажет.

Әрине, ол үшін қазақ тілінің дәстүрлі грамматикаларын формальдық тұрғысынан қайта қарау қажет болады. Себебі мұндай формальды грамматикада жазба және сөйлеу мәтіндеріндегі тілдік бірліктерді формальды жолмен бөлшектеудің, таңбалаудың, мән-мағынасын ашудың және т.б. жан-жақты тұстарының ескерілуін қажет етеді.

Бұл айтылғандар тілдік бірліктердің тұрпаты (формасы) мен мән-мағынасы арасындағы қатынастан, яғни сөз формасы мен мағынасы, сөйлем түрі мен мән-мағынасы және мәтін мен оның мазмұны ара қатынастарынан байқалады.

Қазақ мәтінін автоматты түрде талдау мен жинақтау жүйелерін құрудағы мұндай бірізді шешім таппайтын сипаттамалардың сырын ашу болашақ зерттеулердің міндеті деуге болады.

Әрине, қазақ тілінің грамматикалық құрылымын дәстүрлі әдіспен зерттеу нәтижелері мәтін лингвистикасы саласындағы ізденістерге негіз болары сөзсіз. Осыған орай, қазақ филология саласының тұңғыш профессоры, қазақ тілінің теориялық жағының негізін қалаушы Құдайберген Қуанұлы Жұбановтың «Жаңа грамматиканың жаңалықтары жайынан» атты еңбегінде сөз болған жайттар аталған зерттеулерге арқау болар деген оймен, ондағы тілдік бірліктер мағынасы мен формасына қатысты кейбір тұстарына тоқталайық.

Автор сөзімін айтқанда: «…Грамматиканы тек мағына жүйесіне негіздесе де немесе тек форма жүйесіне негіздесе де, тілдің табиғатын көрсете алмайтын сыңар езулік болады. … Тілдің өсуі ала-құла келеді. Бір жағы ілгері, бір жағы кейін жатады да, сөздің мағынасы мен формасының арасы да үздік-созық келеді. Сондықтан жалғыз мағынаға не жалғыз формаға сүйенген өлшеу, тіпті форма мен мағынаның екеуін қабаттастырып өлшеушілік – мұның қай-қайсысы да – тілді өлшеуге жарамай жүреді…», – дейді [6, 221]. Әрі қарай осы кітаптың 241-бетінде проф. Қ.Жұбанов өз ойын былайша тұжырымдайды: «Грамматиканың мақсаты жалғыз-ақ сөздің қандай мағыналары болатындығын баяндау ғана емес, қандай формаларға қандай мағыналар сәйкес келетінін баяндап, дұрысы, қандай мағына шығу үшін сөзді қалайша құрастыру керек екендігін баяндау. Сондықтан құрастыру, құрылыс жағын елемейтін грамматика грамматика да болмақ емес», – дейді.

Бұл тұжырымдар қазіргі кездегі «Құрылымдық грамматика», «Туындау грамматика» мен «Функционалды грамматика» және «Мәтін лингвистикасы» салаларында жиі айтылып жүрген мәселелер.

Түркі тілдерінің синтаксисіне қатысты зерттеулерде күрделі сөйлемдердің парадигмасына күрделі синтаксистік тұтастық, абзац және төл сөз бен төлеу сөздері бар мәтін үзінділерін қосу қажеттігі де ұсыныс ретінде айтылып жүр [7].

Қазақ тіліндегі мәтін мазмұнын компьютер арқылы түсінуге мүмкіндік тудыратын модель жасау үшін мәтін құрамындағы негізгі тілдік бірліктердің (морфема, сөз, сөз тіркесі, сөйлем, абзац және т.б.) мағына-мазмұндарын формальды түрде анықтауға бағышталған арнайы зерттеулер жүргізілуі қажет деп санаймыз.

Неше түрлі хабарламалардың, ғылыми-техникалық әдебиеттер мен мектеп және жоғары оқу орындарына арналған оқулықтардың қазақ тілінде жарық көріп, сондай-ақ іс қағаздарының мемлекеттік тілде жүргізіле бастағаны бәрімізге мәлім. Олай болса, қазақ тіліндегі мәтіндерді компьютер көмегімен автоматты түрде өңдеп, қажеттілікке сай пайдалану ісі тіл біліміндегі жаңа сала – мәтін теориясымен тығыз шұғылдануды керек етеді. Қарқынды дами бастаған мәтін теориясы көптеген күрделі проблемаларды кеңінен қарастыруға жол ашты.

Ұлттық тіліміз үшін мәтіннің негізгі бірліктерінің мәнін ашып, олардың өзара байланысының сыр-сипатын ашу мәселелері аса маңызды. Сондықтан да қазақ мәтіндерінің ішкі заңдылықтарын егжей-тегжейіне дейін зерттеу – заман талабы деуге болады.

Соңғы кездегі мәтін лингвитикасының нәтижелеріне сүйенсек, мәтін бойындағы абзацтарды «мағыналық квант» немесе «дайын күйдегі» синтаксистік үлгі ретінде түсінік табатын мәтіннің ең кіші логика-семантикалық компоненті (бөлігі) деп қарастыруға әбден болады.

Үндіеуропа тілдеріндегі зерттеулерді негізге ала отрып, өз зерттеулерімізде қазақ тіліндегі мәтіндердің стильдік айырымдары бойынша ажыратылған абзацтарға тән функционалды-мағыналық топтарының статистикалық ерекшеліктерін айқындауды мақсат еттік [8]. Осыған орай, мәтін бойындағы абзацтарды – *баяндау абзацы*, *суреттеу абзацы* және *пайымдау абзацы* деп саралап, үшеуі де қатысты абзацтарды *аралас типті* абзацтар деп атадық. Мұндай *баяндау, суреттеу, пайымдау* типтерінің айырымдық ерекшеліктері мәтін ішінде ең кіші семантика-синтаксистік бірлік, яғни абзац түрінде көрініс табады.

Мәселен, зерттеуге алынған мәтіндердегі ең жиі қолданыстағы *пайымдау* түрлері мыналар: 1) есім түріндегі бағалау, пайымдау түрі; 2) тұрақтылық күй мағынасындағы пайымдаулар; 3) нақты немесе жорамал іс-әрекетті негіздеу мағынасындағы пайымдаулар; 4) кейбір іс-әрекеттің қажеттігін, мүмкіндігін білдіретін немесе қалау не тілеу мағынасындағы пайымдаулар; 5) алдын-ала келісілген іс-әрекет мағынасындағы пайымдаулар; 6) кесімді түрдегі терістеу не тұжырымдау мағынасындағы пайымдаулар.

Кез келген мәтінде (ғылыми стильден басқа) сөйлеудің диалог түрі кездеседі, сондықтан сөйлеуге қатысушы екі жақтың да айтылымдарын (төлеу сөзден басқа) абзацқа жатқызуды ұйғардық. Жоғарыда аталған – баяндау, суреттеу, пайымдау абзацтары түрімен қатар мынадай функционалды-мағыналық түрлер ажыратылды: а) хабарлау мәнді абзацтар; ә) сұрақ мәнді абзацтар; б) әсер ету (лепті мағыналы) абзацтар; в) ниет, ойды білдіретін абзацтар; г) пікір «тудыру» абзацтар; д) аралас абзацтар [8, 118-119].

Сонымен қатар, абзацтар қызметіне олардың орын тәртібіне қарай: мәтіннің бас жағында орналасқан – кіріспе абзацтар, ішкі бөлігінде кездесетін – медиальды абзацтар және мәтіндегі оқиға желісін аяқтайтын – қорытынды абзацтардың статистикасы анықталды [8, 119]. Зерттеу нысаны ретінде үш түрлі стиль бойынша: газет мәтіндері, ғылыми-көпшілікке арналған мәтіндер және көркем әдебиет мәтіндері қарастырылып, оларда кездескен абзацтардың түрлері топталып, статистикалық деректерге компьютер көмегімен салыстырмалы талдау жүргізілді. Алынған статистикалық нәтижелер арнайы кесте түрінде көрініс тапты [8]. Абзацтардың функционалды-мағыналық топтарын нақтылау тұрғысынан абзац түрлерінің мазмұндық жағынан заттық-логикалық топталуы да мәтінді «туындату» мәселесінде маңызды орын алады.

Сонымен, «Қазақ тілінің мәтін туындату алгоритмін» құру үшін ең алдымен мынадай мәселелер шешім табуы қажет:

1) формальды түрде мәтіннің жалпы тақырыбының берілуі;

2) мәтіннің негізгі семантика-синтаксистік бірліктерінің айқындалуы;

3) аталған бірліктерді тұтастықта ұстайтын ықтималды-анық ережелердің берілуі.

Қазақ мәтіндерінде аталған проблемалық мәселелердің ойдағыдай шешілуі, әрине, әртүрлі құрылым-көлемдегі мәтін түрлері мен типтерін жаңа технологиялық әдістермен терең зерттеуді қажет етеді.

Әдебиет

1. *Новиков Э.А., Егоров В.С.* Информация и исследователь. – Л.: Наука, 1974.

2. *Алейник А.* От интеграла до рубля // Молодой коммунист. –1974. -№5. – С. 50-57.

3. *Марчук Ю.Н.* Проблемы машинного перевода. –М.: Наука, 1983.

4. *Гальперин И.Р.* Текст как объект лингвистических исследований. – М.: Наука, 1981.

5. *Кузнецов И.П.* Некоторые механизмы порождения текстов описания // НТИ. Сер. 2. –1975. -№9 – С. 10-15.

6. *Құдайберген Жұбанов.* Қазақ тілі жөніндегі зерттеулер. –Алматы, «Ғылым», 1999. – 582 б.

7. *Закиев М.З.* Актуальные проблемы в сложных предложениях в тюркских языках. – Баку, 1983.

8. *Аскар Кудайбергенулы Жубанов.* Основные принципы формализации содержания казахского текста. – Алматы, 2002. – 250 с.

ОҚУ ЛЕКСИКОГРАФИЯСЫНЫҢ ҒЫЛЫМИ ПӘН РЕТІНДЕГІ ЕРЕКШЕЛІГІ

ҚР ҰҒА «Хабарлары», филология сериясы.
Алматы, 2006, №5, 20-25 бб.

Егемен еліміздің орта және жоғары оқу орындарында салалық пәндерді мемлекеттік тілде оқыту көптеген күрделі, кезек күттірмейтін мәселелерді шешуді талап етіп отыр. Осыған байланысты басты мәселенің бірі – сөздіктер, оқулықтар мен оқу құралдарын құрастырғанда педагогикалық процеске негіз болтын қазақ тілінің жалпы лексикалық қорын және салалық пәндердің терминологиясын құрастыру.

Қазақ тіліндегі әр салалық пәндерге қатысты терминдерін қалыптастыруда орыс халқы мен туыстас түркі халықтарының озық тәжірибелеріне сүйенетін ғалым-лексикографтарды күрделі жұмыстар күтіп тұр деуге болады. Атап айтар болсақ, халық ауыз әдебиетінің бай мұрасы мен

қазақ халқының тұрмыс-тіршілігіне, салтына тән көптеген ежелгі атаулар мен терминдерді оқырманның жас ерекшелігіне, білім деңгейіне қарай сөздіктер мен оқу құралдарында орынды пайдалана білу жүзеге асыр еді.

Қай ғылым саласын алмасақ та, оның өзіне тән метатілі болатыны мәлім. Метатіл ұғымына тек терминдер ғана емес, жалпы ғылыми лексика да енеді. Ал мұндай метатілдің қалыптасуына сөздіктердің де әсері болатындықтан, көпшілік ғалымдар тіл білімінің лексикография саласын теориялық және практикалық негіздемесі бар ғылыми пән ретінде қарастыруды жөн санайды.

Ахмет Байтұрсынұлы мен Құдайберген Жұбанов бастап дамытқан қазақ тіліне қатысты терминдер белгілі дәрежеде қалыптасқанымен, басқа ғылыми пәндердің қазақша терминдік жүйесі әлі де жинақтауды, зерделеуді қажет ететіні белгілі. Сондықтан да оқулықтарды, сөздіктерді пайдалана отырып, олардың түсініктемесін, түсіндірілуін айқындау қажет болып отыр.

Қазіргі кезеңде қазақ тілінің сөздік қоры жаңарып, тіл білімінің арнасы кеңейуіне байланысты оқулықтар мен сөздіктер де уақытқа сай өзгеріске ұшырауда. Жарық көріп жатқан сөздіктерде соңғы жылдары пайда болып жатқан қазақ және орыс тілдеріндегі жаңа сөздер мен байырғы сөздердің түсіндірмелері мен қолданыстары енгізілуде. Сол сияқты ондай сөздіктерде қазақ әдеби тілінің нормаларына сәйкес және күнделікті қарым-қатынаста, оқыту жүйесінде жиі қолданылатын сөздер орын алуда. Айтылуы қиын қазақ немесе шет тілінің сөздері мен сөз тіркестерінің транскрипциясы беріліп, реестр сөздерінің стилдік ерекшеліктері қалыптасқан ескертпелер мен шартты белгілер арқылы көрсетіп беру жұмысы да атқарылып жатыр. Әр түрлі мақсатты ұстаным ететін сөздіктерде сөздің қолдану сипатын жан-жақты қамту үшін кеңінен таралған қазақтың мақал-мәтілдері, сөздердің негізгі және ауыспалы мағыналарын айқындайтын сөз тіркестері келтіріліп, омонимдер мен көпмағыналы сөздер қатарларына ерекше көңіл бөлу ойластырылуда.

Әрине, қазақ тілінің лексикография саласының дамуына тіл білімінің төрінен орын алған мынадай іргелі еңбектер негіз болғаны мәлім: «Қазақ тілінің түсіндірме сөздігі» (1-10 том, А., 1974-86), І.Кеңесбаевтың «Қазақ тілінің фразеологиялық сөздігі» (А., 1977), Н.Т.Сауранбаевтың жетекшілігімен жарыққа шыққан «Казахско-русский словарь» (М., 1954), Р.Сыздықтың жетекшілігімен жарық көрген «Жаңа атаулар» (А., 1999), А.Байтұрсынұлы атындағы Тіл білімі институты лексикографтарының құрастыруымен шыққан «Қазақ тілінің сөздігі» (А., 1999), «Қазақша-орысша сөздік» (А., 2001), «Орысша-қазақша сөздік» (А., 2005) және т.б.

Қазақстан Республикасы Үкіметінің жанындағы Мемлекеттік терминологиялық комиссиясының әр жылдары ресми бекіткен түрлі сала терминдері жинақталып, дүркін-дүркін жеке сөздік түрінде шығарып тұру да дәстүрге айналды. Осындай лексикографиялық жұмыстардың нәтижесі ретінде 1976, 1979, 1990 жылдары тілші ғалым марқұм А.Әбдірахмановтың құрастыруымен шыққан «Атаулар мен терминдер» сөздігі, 1996 жылы жарық көрген Б.Қалиев пен Ш.Құрманбайұлының «Орысша-қазақша және қазақша-

орысша терминдер сөздігі», Тілдерді дамыту департаментінің 1998 жылы Т.Алпысбаев пен А.Әбдірахмановтың және 1999 жылы жарық көрген құжаттар жинағының басын қосып 2002 жылы басылып шыққан «Терминологиялық жинақ», сондай-ақ салалық терминологиялық орысша-қазақша, қазақша-орысша сөздіктердің 31 томы «Рауан» баспасынан жарық көрді (2000ж.). Осы аталғандардың қатарына 1998 жылы «Ғылым» баспасынан шыққан Б.А.Жұбановтың 25677 сөзді қамтитын көптілді (ағылшын, неміс, орыс және қазақ тілдері бойынша) «Химиялық сөздікті», А.Салқынбай мен Е. Абақанның авторлығымен «Сөздік-словарь» баспасынан 1998 жылы шыққан «Лингвистикалық түсіндірме сөздікті», Т.Момынов пен А.Рақышевтың 1999 жылы «Атамұра» баспасынан жарық көрген орысша-қазақша «Медициналық сөздікті» жатқызуға болады. Сол сияқты М.Қазбеков пен Р.Шойбековтың «Русско-казахский словарь» атты сөздіктің (А., 1996), 2002 жылы «Аруна» баспасы шығарған 30000 сөзді қамтитын мектеп оқушылары мен жоғары оқу орындары студенттеріне арналған «Сөздіктің» (10000-нан аса сөз, құрастырғандар: Б.Исмағұлова, Э.Ережепова, Г.Әбдіжапарова) және үстіміздегі 2006 жылы Астанада жарық көрген З.Ө.Жантекеева мен О.Нақысбектовтың «Қазақша-орысша және орысша-қазақша оқушыларға, студенттерге арналған сөздіктің» (14000 сөз) де қазақ және орыс тілдерін оқып-үйрену кезінде аса құнды лексикографиялық еңбектер екені белгілі.

Осы бағыттағы аталған және басқа да аттары аталмай қалған көптеген басылымдардың қазақ тілін оқыту процесінде, әсіресе жоғары білімді, білікті мамандар дайындауда игі көмегін тигізері сөзсіз. Қазіргі жағдайда оқу процесін қамтамасыз ету үшін арнайы маман-педагогтар мен ғалым-лексикографтардың араласуымен қазақ тілінің оқулықтарын, оқу құралдарын, көптілді аудармалы, түсіндірмелі және энциклопедиялық сөздіктерін сапалы құрастыру заман талабы деуге болады. Өкінішке орай, мектеп оқушылары мен студенттерге арналған қолымызда бар бірқатар сөздіктерде, оқулықтар мен оқу құралдарында әлі де болса теориялық тұрғыда кеткен кейбір олқылықтар жоқ емес.

Сондықтан оқырман назарына ұсынып отырған мақаламызда осындай кемшіліктерге болашақта жол бермеу тұрғысынан оқу лексикографиясына қатысты кейбір теориялық және практикалық мәселелерді сөз етуді жөн көрдік.

Оқу лексикографиясы жалпы лексикографияның жас саласы болып саналады. Солай бола тұра, оның да өзіне тән жақсы зерттелген тарихы мен дәстүрі бар. Ресейде ең алғашқы оқу лексикографиясының орталығы тек 1966 жылы ғана құрылып, Мәскеудегі А.С.Пушкин атындағы орыс тілі Институтында сектор ретінде қызмет етеді. Қазіргі кезде аталған сектор орыс тілінің біртілдік және екітілдік (орысша-шеттілдік) сөздіктерін құрастырудың теориялық және практикалық жұмыстарын кең көлемде атқаруда.

Оқу лексикографиясының ғылыми пән ретіндегі өзіне тән ерекшелігі, оның жалпы лексикография, лексикология, семасиология психо- және әлеуметтік лингвистика және т.б. салалардың тілді үйрету әдістемесіне

қатысты мәселелерімен тығыз байланысты болуында. Сондықтан оқу лексикографиясы дәстүрлі жалпы лексикографияның мәселелерімен қатар, мынандай көптеген әдістемелік факторларды ескеруді қажет етеді: оқыту кезеңдеріне қатысты сөйлеу әрекетінің түрі, тілді үйренушілер үшін тілдің ана тілі (немесе аралық тіл) екендігі және қажетті тілдік ортаның болу-болмауы, жас ерекшелігі мен жалпы мәдени деңгейге қатысты болуы және т.б. мәселелер.

Осы аталған лингвистикалық және методологиялық факторлар тек бірдей (және кешенді) дәрежеде ескерілсе ғана оқу лексикографиясын пайдалануға байланысты, ана тілін не шет тілдерін үйретудегі оқу сөздіктеріне қатысты дұрыс әдістемелік нұсқау беруге болады. Лексикографияның жалпы теориясы принциптерін орынды пайдалану негізінде тілді үйрету жолындағы ескірген әдістерге жол бергізбей, ішкітілдік және тіларалық баламалардың шегін анықтауға, лексика-семантикалық тіркесім және т.б. айқындауға мүмкіндік туындайды.

Әдістемелік факторларды ескеру негізінде мына жайттар анықталады [Марусенко, 1996, 309-б.]: 1) сөздіктің адресаты (мектеп оқушысы, студент және т.б.); 2) сөздіктің типі (біртілдік немесе аударма); 3) сөзтізбе көлемі (минимум сөздік, орта немесе толық); 4) сөздік мақаласының құрылымы, ондағы лексикалық материалдың берілу әдіс-тәсілі (эквиваленция, дефиниция, безендірілу, лексика-семантикалық топтағы басқа сөздерге қарама-қарсылығы) синонимия, анонимия, паронимия, ассоциативтік, тіркесімдік, грамматикалық, орфоэпиялық және стильдік ерекшеліктері және т.б. 5) полиграфиялық орындалуы.

Лексикорграфияның жалпы теориясын дұрыс ескеру тілдік ақиқаттықты шынайы көрсете білуге мүмкіндік жасайды. Ал методикалық талаптар оқу сөздігіндегі ең маңызды деген жайттарды, мысалы, сөздіктің тілді үйренуге бағытталуын, белгілі оқыту кезеңі мен оқыту формасын қолдануын, яғни оның қызметін анықтайды.

Оқу лексикографиясының тәуелсіз лингводидактикалық пән болуының мәні, оның тіл білімі мен методиканың өзара тығыз байланыстылығы мен өзара қатыстылығында жатыр.

Шынында да, егер қазіргі түсінікте, оқу лексикографиясы тіл білімінің ең бір жас саласы бола тұра, практикалық тұрғыдан алғанда оны бұрыннан келе жатқан сала деуге болады. Бұлай тұжырымдау үшін, кез келген сөздікті, ең алдымен, дидактикалық шығарма деп түсінген жөн. Лексикограф, яғни сөздік авторы оқулықтың авторы тәріздес, қоғам (ұжымдық тіл иесі) мен жеке адам аралығындағы делдал ретінде қызмет етеді. Сондықтан кез келген сөздік типінің қызметі – жеке адамның санасындағы білім мен ұжым санасындағы білім аралығындағы айырмашылықты жоюға талпыныс деуге де болады.

Қолданбалы тіл білімінің ірі маманы В.Г.Гак сөздіктердің қызметін мынадай негізгі төрт топқа бөліп қарастырады:

1. Тіларалық қатынасты ұйымдастыруға байланысты. Мұндай қызмет көбінде екітілді сөздіктер мен тілдескіштер арқылы іске асады.

2. Тілді оқыту (үйрету) үшін және мағынасы қиын сөздерге түсініктеме беру арқылы.

Сөздіктің мұндай қызметінде лексикографиялық материалдар ретінде біртілді немесе екітілді сөздіктер, глоссариялар, тілдескіштер қолданылуы мүмкін. Оқыту процесі мынадай мәселелерге қатысты болады: а) сөйлеу тілінен айырымы бар көне мәдениет тіліне қатысты тілдерді меңгеруде. Мысалы, ол ежелгі дәуірге қатысты тіл немесе басқа да тілдерге – латын, ежелгі грек тілі, Европадағы орта ғасырлық ескі словян тілі, Шығыс елдеріндегі классикалық араб тілі және т.б. тілдер болуы мүмкін; ә) шет тілдеріне; б) ана тіліне.

3. Ана тілін сипаттау мен нормалауға байланысты. Мұндай қызметтер, ең алдымен, түсіндірме сөздіктер арқылы және мына сияқты арнаулы біртілдік сөздіктер арқылы арқылы іске асады: орфографиялық, қиын сөздер сөздігі және т.б.арнаулы біртілдік сөздіктер арқылы іске асады.

4. Тілдерді зерттеу ісіне байланысты. Мұның негізінде біртілді және екітілді филологиялық сөздіктердің әр түрлі типтері құрастырылады [Гак, 1977].

Лексикографияның аталған қызметтерін тарихи тұрғыда қарастырсақ, оның даму сатыларын былай жіктеуге болады:

1. Сөздікке дейінгі кезең (Еуропада – XV ғ. дейінгі кезең). Бұл кезеңде сөздік нәтижелері глоссариялар, вокабулалар (хрестоматияға не оқулық тарауына қатысты берілетін қысқаша сөздік), ономастикондар түрінде көрініс тапты. Олардың негізгі мақсаты мәдениет тілінің қиын сөздерінің мағынасын ашу.

2. Сөздік түзудің алғашқы кезеңі (Еуропада – XV-XVII ғасырлар). Бұл кезеңде лексикондар (белгілі қызмет аясына ғана тән сөз қоры) мен тіларалық қарым-қатынас және мәдениет тілін зерттеу мен шет тілін оқыту қажеттігі үшін қызмет ететін екітілдік сөздіктер өмірге келді.

3. Лексикографияның дамыған кезеңі (Еуропада – XVIII ғ. бастап) Бұл кезең сөздіктердің әр түрлі типтерінің, әсіресе түсіндірме сөздіктердің шыға бастауымен ерекшеленеді [Марусенко, 1996, 310-б.]. Құрастырылған сөздіктер лексикография саласының жоғарыда аталып өткен төрт түрлі қызметтерін іске асыруды мақсат етеді.

Сонымен, оқу лексикографиясы сөздік құрастырудың алғашқы даму кезеңінің өзінде-ақ пайда болды деуге болады. Себебі бастапқы кезеңдердегі сөздіктердің аттарының өзі-ақ, олардың дидактикалық міндеттерін нұсқап отырды. Мәселен, бірінші латынша-ағылшынша сөздік «Promptulorium parvulorum» деп аталған, яғни «Балаларға арналған сөз қойма». Ертеректегі французша-латынша сөздіктер – «Commentarius puerorum», яғни «Жастарға арналған түсіндірме сөздік» деп және т.б. аталған. Бері келе мұндай сөздіктердің оқуға бағыттағыш сипаты екінші орынға ауысып, біріншіге – түсіндірме, тарихи, этимологиялық сөздіктер типтеріне ауыса бастайды.

Лексиканы оқыту мақсатының нысаны етіп алуға арналған әдебиеттерді талдай келе, ғалым В.В.Морковкин мынандай негізгі бес аспектіні бөліп қарастырады [Актуальные проблемы …, 31-35-бб.]:

1. **Оқу сөздіктері**. Мұндай сөздіктерге қатысты мәселелер, әсіресе сөзтізбе құрылымын талдау мен оқу сөздіктерінің типтері жайлы мәселелер толық зерттелді деуге болады. Бірақ бұл зерттеулердің негізгі мақсаты әр типті сөздіктерде орын алатын ұқсастық пен айырмашылықтардың аналитикалық мәселелері, лексиканың семантикасын анықтаудың тәсілдері және т.б.

2. **Лексикалық минимум сөздіктері**. Қазіргі кезде тілдің сөздік қорын минимизациялау теориялық жағынан барынша зерттелді деуге болады. Соның нәтижесінде оқу формасына қарай көптеген лексикалық минимум сөздіктері құрастырылды. Бірақ оқу процесінде лексикалық минимумның барлық артықшылығын орынды пайдалану үшін, ол жан-жақты сипатта жүзеге асуы және оқу барысының барлық кезеңдерін түгел қамтуы қажет. Әрине, бұл аталған мәселелердің әлі де толық шешімі табылмай келеді.

3. **Лингвостатистикалық зерттеулер**. Мұндай зерттеулердің нәтижелері әдетте жиілік сөздіктер мен үлестірімді жиілік сөздіктер (қолдану жиілігі әр стиль бойынша үлестіріліп берілген) түрлерінде көрініс табады.

4. **Лексикографиялық типтегі лексикаға қатысты оқу құралдары**. Әр түрлі тілдерге арналған мұндай оқу құралдары көптеп шығарылды деуге болады. Бірақ олар, теориялық тұрғыда, әлі де болса зерттеу нысаны бола алмай отыр.

5. **Оқулықтар**. Мұндай оқулықтарда лексиканың семантикалық жақтары әр тақырыпқа қатысты сабақтық сөздіктерде, жинақтық (сводный) сөздіктерде және арнайы лексикалық жаттығуларда қарастырылады. Қандайда болсын оқулықтар үшін лексикалық аспект өте-мөте маңызды деп саналады. Бірақ мұндай оқулықтарда лексиканы семантикалау тәжірибесінің теориялық негіздемесі орын ала бермейді.

Жоғарыда көрсетілгендей, оқу сөздігі, ең алдымен, қажетті әдістемелік бағытталуымен, нақты оқу кезеңдеріне қатыстылығымен, оқудың мақсаты мен міндетіне және т.б. қарай сипатталады. Оқу сөздіктері тілдік материалды таңдау ісінде арнайы мақсат қоюмен ерекшеленеді. Яғни болашақ лексикографиялық талдау үшін қажетті болатын тірек бірліктердің негізгі тізімін түзу ісімен ерекшеленеді. Сөздіктің сөзтізбесі әрі шағын, әрі жеткілікті сипатта болуы қажет. Бұл мақсат тек сөздің негізгі мағынасындағы кавантитативті параметрлерін, грамматикалық тұрпатын және синтаксистік құрылымын, тақырыпты ашатын ең маңызды сөздер мен сөз тіркестерді таңдауды қатаң түрде есепке алу арқылы ғана жүзеге асады. Осымен бірге сөздіктің негізгі тақырыбы мен сөйлеу жағдаятын қамтуды тексеру қажеттігі де туындайды. Өкінішке орай, тәжірибе жүзінде мұндай тексеру жұмысы көбінде орын ала бермейді.

Оқу сөздігінің сөзтізбе көлемі, ең бірінші, оның типіне қарай анықталады. Мысалы, қысқаша түсіндірме сөздіктегі лексика осы тілдік ұжымның әр мүшесі үшін қалыпқа айналған сипатта болуы қажет. Аталған мәселені шешу үшін мәдениеттің және ғылым мен өндірістік білімнің орта жиынтығы қамтылған ең кіші сөйлеу қызметінің моделі құрастырылуы керек. Мұндай түсіндірме сөздік-анықтағыш 20000-25000 сөздік мақаласынан

тұратыны белгілі. Бірақ мұндай сөздік пассивті немесе жеке индивидтердің активті сөздік қорының шекаралығында тұрады да, оқу сөздігінің активті типіне жатады.

Кез келген оқу сөздігі қатаң мақсаттылықпен жүргізілген таңдаудың нәтижесі, яғни кең түсініктегі лексикалық минимум бола тұра, ол ішкі диалектикалық қарама-қарсылыққа толы. Бір жағынан, ондай сөздік, белгілі дәрежеде, нақты бағыттағы мәтіндерді және оқыту кезеңіне қатысты кейбір қиындықтарды қамтуы қажет. Ал екіншіден, ондай сөздіктерде кез келген мәтінді түсінуге мүмкіндік туғызатын нақты материалдар да орын алуы қажет. Бір ғана принципті (жиіліктік, тақырыптық, ситуативтік және т.б.) ұстаным ететін сөздік типі қойылған талапты орындай алмайды.

Сөздіктің әр деңгейіне тән сөздер бірнеше топ құрайды. Оларға жататындар [Марусенко, 1996, 313-б.]: а) тілдің грамматикалық қаңқасын құрайтын және кез келген мәтінде кең қолданыс табатын – құрылымдық сөздер; ә) көптеген мәтіндерде қолданылатын және сол себепті нақты тақырыпқа, сөйленіс жағдаятына байланыссыз келетін – жиілігі жоғары негіз сөздер; б) белгілі тақырыппен немесе жағдаятпен тығыз байланысты болатын – тақырыптық, жағдаяттық сөздер. Бұл категорияға қатысты сөздер жоғары жиілікке ие емес, шектеулі сөйлеу мәнмәтіндерінде қолданылады. Сөздікке тек әр түрлі категориядағы (құрылымдық, тақырыптық, жағдаяттық) бірліктерді енгізу ғана кең шеңбердегі мәтіндерді түсінуге мүмкіндік туғызады.

Оқу сөздігінің құрылымдық бөліктерінің өзгешелігі. «Үлкен» (ірі) сөздіктерге қарағанда оқу сөздіктерінің құрылымы негізгі екі ерекшелікпен сипатталады: лексикалық бірліктерге түсініктеме берумен (семантикалау) және ол бірліктердің парадигматикалық, синтагматикалық қасиеттері мен қатынастарын анықтайтын әдіс-тәсілдерімен.

Лексикалық бірліктерді семантикалау түсіндірме оқу сөздігінің маңызды құрылымды бөлігі болып табылады. Әрине мұндай компонент белгілі дәрежеде басқа да сөздіктер типіне де қатысты. Аударма оқу сөздіктерінде семантикалау, әдетте, аударма баламалары арқылы, ал түсіндірме сөздіктерде мұндай семантикалау – анықтамалар арқылы іске асады. Оқу сөздіктерінің маңызды құрылымдық бөлігі саналатын анықтама: шағындық, көрнекілік және қарапайым сөзге саятындай талаптарға сай келетіндей болуы керек.

Шетел азаматтарына бағышталған түсіндірме сөздіктерде қайсыбір ұғымға түсініктеме беретін анықтамаларында таныс сөздердің шағын көлемді бөлігін пайдаланған жөн болады. Бұл сөздер «минимум сөздік» құрайды және осылардың негізінде сөздіктің басқа да лексикалық бірліктері анықталады. Мысалы, кең танымал М.Уэстің [West, 1965] сөздігінде 24000 сөздің мағынасы көлемі 1500 сөздерге тең лексикалық топ мағыналары арқылы анықталады. Сол сияқты француз тілінің негізгі лексикалық тобын қамтитын [Gougenheim, 1959] анықтағышта 55 сөз арқылы және қарапайым сөздіктің 1374 сөздері бойынша басқа сөздердің мағыналық анықтамалары берілген.

Оқу сөздіктерін құрастыру тәжірибесінен байқалатын жайт, ол «минимум сөздіктер» негізінде түзілген сөздіктер, көбінде, жасандылық пен нақтылықтың жеткіліксіздік сипатымен өзгешеленеді. Бірақ мұндай олқылықтар семантикалау әдісінің кемшілігінен емес, құрастырушылардың жіберген қателігінен деуге болады.

Оқу сөздіктері үшін аса маңызды мәселе сөздердің синтагматикалық және парадигматикалық байланысын бейнелеп көрсете білу. Үйренуге тиісті тілдің тек лексикасын ғана меңгеру сөйлеу тіліндегі оның дұрыс қолданылуына кепіл бола алмайтындығы белгілі. Ана тілі мен шет тілінің түйісу әсерінен үйренушінің тілінде бұрын-соңды қолданылмайтын кейбір сөз тіркестерін туындатады. Ірі филологиялық сөздіктерде сөзқолданыстардың ерекшеліктері жайлы, яғни сөздердің тіркесімдігі жайлы мәліметтер бола бермейді. Сондықтан, тәжірибе жүзінде, осындай мәліметтер бере алатын, сөздердің ең ықтимал нормативті байланысын бейнелейтін оқу сөздіктері қажет-ақ [Регинина и др., 1980].

«Парадигма» дегеніміз бір-біріне қарама-қарсы қойылған, бірақ ортақ белгісі бар немесе ұқсас ассоциация тудырушы кез келген лингвистикалық бірліктер тобы екені белгілі. Міне осындай парадигматикалық қатынастарды бейнелейтін сөздіктерде тақырыптық, синонимдік, антонимдік топтар көрініс табады.

Оқу сөздіктерінің типологиясы. Типологиялық тұрғыда оқу сөздіктері бірнеше көзқараста қарастырылады: 1) сөздіктің қызметі тұрғысынан; 2) сөздіктің экстралингвистикалық параметрі бойынша; 3) сөздіктің ішкі лингвистикалық параметрі бойынша.

Оқу сөздіктері қызметіне қарай, жалпылама және мамандандырылған сөздіктер деп те ажыратылады. Бірінші сөздік үйренушілердің барлық категорияларына арналады да, ал екіншісі үйренушілердің тек белгілі категорияларына ғана арналады. Мысалы, мектеп оқушыларына, филолог-студенттерге, филолог емес студенттерге, мұғалімдерге және т.б.

Сөздіктің экстралингвистикалық параметрлері әрбір сөздіктің өзіне тән ерекшелігін анықтайды. Мысалы, олардың энциклопедиялық сөздікке жақындығына қарай немесе оның кейбір элементтерінің ғана энциклопедиялық сөздікке енгендігі бойынша. Осы тұрғыда мынадай бірліктерді бөліп қарастыруға болады: материалдық мәдениеттің заттық сипаты, рухани мәдениеттің концептері, ел тану компоненті және үйренуге (зерттеуге) қажетті тілдегі логикалық категориялардың бейнелену сипаты. Яғни «әлем бейнесінің» әр түрлі жекелік, топтық, ұлттық санасының концептуалды бейнеленуі. Экстралингвистикалық параметрлер жалқы есімдердің, мақал-мәтелдердің, қанатты сөздердің (сөйлемшелер) және т.б. сөздіктері үшін аса маңызды рөл атқарады.

Ішкі лингвистикалық параметрлер негізіндегі сөздіктер типологиясы тілдің сөздік құрамының типологиясын бейнелеуі қажет. Кез келген тілдің сөздік құрамы жүздеп-мыңдап саналады. Сондықтан қажетті сөзді тез іздеп табу үшін лексиканың жүйелік сипатын ескеретін стратегияны туындатып алған жөн болады. Соның негізінде қажетті деген сөзді барлық сөздік

құрамының ішінен іздемей, тек оның кішігірім бөлігінің ішінен, мысалы, семантикалық өріс ішінен немесе лексика-семантикалық топ ішінен және т.б. іздеу мүмкіндігі туындайды. Оқу лексикографиясының теориясы, ең ұтымды жағдайда, тіл жүйесінің лексикалық аспекті бойынша, аспектілік сөздіктер жүйесін құруды қалайды. Сол арқылы іздеудің табиғи стратегиясы модельденіп, қажетті сөзді іздеп-табу мезгілі қысқарады.

Бұл мәселе сөздіктің көлеміне, оларды ұйымдастыруға, сөздіктегі әрбір лексикалық бірлікті сақтайтын ақпарат көлеміне, әр қилы таңбалар жүйесіне, сілтемелер жүйесін құру жұмыстарына қатысы бар. Яғни бұл мәселе кешенді (жан-жақты) сипатқа ие деуге болады.

Қазіргі кезде халықаралық қатынас нығайып, көптеген дербес ұлттық мемлекеттердің пайда болуы және бұрынғы отарлықта қалған халықтардың мәдениетінің көтерілу сипаты, қайтадан лексикографияның тіларалық қарым-қатынасқа және тілдерді оқытуға байланысты қызметтерін бірінші орынға ауыстыруына жағдай жасайды. Әлемдік өндірісті интеграциялау кезеңінде әртүрлі түсініктерді бір ізге түсіру де аса қажет болуда. Сондықтан егемен Қазақстанның халықаралық деңгейге шығуына және бірқатар жетекші елдермен қазіргі таңдағы ғылым мен технологияның әр түрлі салаларында ғылыми және өндірістік ақпараттармен алмасудың едәуір ұлғаюына байланысты, неғұрлым кең ауқымды қамтитын көптілді сөздік қажеттілігі де туындауда. Жақын болашақтағы лексикографиялық зерттеулердің негізінде оқу сөздіктерінің әр ұлттық тілдерге, соның ішінде қазақ тіліне қатысты да практикалық маңызы зор жүйесі құрылуы мүмкін.

Сонымен, мақалада сөз болған және де осы бағытқа қатысты басқа да теориялық жайттар ескерілетін болса, қазақ тілінің оқу лексикографиясына байланысты көптеген әрекеттерді компьютер көмегімен толық автоматтандырып, олардың электронды түрлерін көптеп шығару жүзеге асар еді.

Әдебиет

1. Актуальные проблемы учебной лексикогрфии. М., 1977.
2. Гак В.Г. О некоторых закономерностях развития лексикографии (учебная и общая лексикография в историческом аспекте) // Актуальные проблемы учебный лексикографии. М., 1977.
3. Жантекеева З.Ө., Нақысбеков О. Қазақша-орысша және орысша-қазақша оқушыларға, студенттерге арналған сөздік. – Астана, 2006. – 208 б.
4. Жұбанов Б.А. Химиялық сөздік ағылшын, неміс, орыс және қазақ тілдерінде. – Алматы: «Ғылым», 1988. – 200 б.
5. Исмағұлова Б., Ережепова Э., Әбдіжапарова Г. Оқушыларға, студенттерге арналған қазақша-орысша, орысша-қазақша сөздік. – Алматы: «Аруна», 2002. – 416 б.
6. Марусенко М.А. Учебная лексикография // Прикладное языкознание: Учебник / Отв. Редактор А.С.Герд. СПб.: Изд-во С.-Петербург. Ун-та, 1996. – 528 с.

7. Момынов Т., Рақышев А. Медициналық сөздік (орысша-қазақша). – Алматы: «Атамұра», 1999. - 312 б.

8. Регинина К.Б., Тюрина Г.П., Широкова Л.И. Устойчивые словосочетания русского языка. М., 1980.

9. Салқынбай А., Абақан Е. Лингвистикалық түсіндірме сөздік. – Алматы: Сөздік-Словарь, 1998. – 304 б.

10. Gougenheim G. Dictionnaire fondamemtal de la langue francaise. Paris, 1959.

11. West M. An International Reader's Dictionary. London, 1965.

АХМЕТ ЖҰБАНОВ ЗЕРТТЕУЛЕРІНДЕГІ ҚАЗАҚТЫҢ КҮЙ ЖАНРЫ МЕН ДОМБЫРА МУЗЫКАЛЫҚ АСПАБЫ ТУРАЛЫ

Ақтөбе облыстық әкімдігі. Ақтөбе облыстық мәдениет басқармасы. Облыстық С.Бәйішев атындағы әмбебап-ғылыми кітапханасы. «Ғасырлар пернесінің жыршысы» атты академик, композитор А.Қ.Жұбановтың 100-жылдығына арналған республикалық ғылыми конференция материалдары. – Ақтөбе, 2006, 114 бет. (Мақала: 43-49-бб.)

Әрбір халықтың рухани-мәдени өміріндегі ең көне әрі ең маңызды құбылыстарының бірі музыка, ән-күй өнері. Ұлттық мәдениетті, оның үлкен бір саласы музыкалық мәдениетті қолдау, дамыту, бүгінгі таңда мемлекеттік іске айналып отыр.

Ғасырлар бойы қалыптасқан және бүгінге жетіп, қалың жұртшылық игілігіне айналып отырған қазақ музыка аспаптарының құнарлы табиғаты мен бай қорына зерделі көзбен қарайтын болсақ, бұл өнер саласы, яғни музыка мен музыка аспаптары – қазақтың өз топырағында дамыған өз даналығымен дүниеге келген өнер, өз қабілет мүмкіндігімен өрбіген мұра, сирек кездесетін құбылыс (феномен), абзал құндылық.

Ғалым Жиенбек Рсалдиннің «Ахмет Жұбанов» атты кітабында былайша пайымдалған: «Музыка тарихшысы Ахмет Жұбановтың еңбектерінде халық музыкасы, оның өмірі, күн көріс кәсібімен тығыз байланыста сөз болады. Қазақ халқының басынан өткізген неше түрлі тарихи оқиғаларының бәрі де оның ән-күйінен, музыкалық шығармашылығынан орын алады. Халықтың қуанышы мен қайғысы, арманы мен азабы, елдігі мен тәуелсіздік жолындағы ерлік күресі, міне, осының бәрі де оның ән-күй шығармашылығынан көрініс тапқан. Қазақ халқының рухани байлығы болған ән-күйі – қашан да оның өмірінің ұмытылмас шежіресі, ұрпақтан ұрпаққа жететін өшпес үні екенін осылайша сөз етеді. Қазақ музыкасындағы негізгі жанрдың бірі – «Күйдің» этимологиясын, оның музыкалық шығарма ретінде ерекшелігін, құрылысын, шығу тарихын алғаш зерттеген ғалым – Ахмет Жұбанов» - деп тұжырымдайды.

1964 жылы Москвада өткен антропология және этнография

ғылымдарының жетінші халықаралық конгресінде жасаған «Қазақтың музыкалық халық аспабы – домбра» деген тақырыптағы баяндамасында академик Ахмет Жұбанов мынадай тұжырымдар жасайды [1, 6-7-бб.]: «Домбырашы-солистердің репертуары домбыраға арналған «күй» - деп аталатын аспаптық пьесалардан тұрады. «Күй» – деген сөз XI ғасырда-ақ белгілі болған. Махмуд Қашғаридің «Дивани-луғати-түркі» атты еңбегінде «көк» деген сөз (қазақша айтылуында – «күй») аспаптық музыка мен әндер деген ұғымды бірдей қамтиды. Тіпті қазірдің өзінде татарлар әнді де, аспаптық музыканы да «күй» дейді. XIV ғасырдан бастап «күй» деген сөз аспаптық музыканың ұғымын білдіреді. Мысалы, Абдықадыр Мурагидің «Зубдатал – адвар» атты еңбегінде аспаптың музыка мен вокальдық музыка дараланады: ***«көк а»*** яғни ***күй*** – аспаптық музыка, ***«дола ыр»*** деген сөз ***«ән»*** деген ұғымды білдіреді» - дейді.

Сонымен бірге Ахмет Жұбанов: Қазақ халқының өмір сүру жағдаятымен тығыз байланыста болған ***күй*** жанрының ғасырлар бойы дамып отырғанын, мұның музыкалық фольклорымызда жүздеп саналатынын, нота жазу мәдениеті жоқ болғандықтан күйлердің біздің заманға ***«ауыз дәстүрлік»*** түрінде жеткенін ғылыми тұрғыда сөз ете отырып, «күйлер сан жағынан да, сапа жағынан да халқымыз мақтан етуге тұрарлық баға жетпес асыл қазына», - деп қортындылайды [3, 91-б.].

Ахмет Жұбановтың көрсетуі бойынша халық музыкасында эпикалық күйлер де, лирикалық күйлер де, тұрмыс-салт күйлер де бар. Ал батырлық рухтағы күйлер сипаты жағынан эпикалық күйлерге жақын келеді. Халық күйлерінде, әсіресе лирика-философиялық жанрға жататын күйлер өлімді, дүлей қара күшті жеңіп шыққан өнерпаздықты, халықтың оптимистік ой-арманын жыр етеді, - деген ойға меңзейді (мысалы, «Қорқыт күйі»). Ахмет Жұбанов күйлер қоғамдық ой-сананың музыка әуеніндегі тікелей көрініс екенін, күй - дәуір үні, замана сыры екенін осылайша дәлелдей отырып, күйдің құрылысы, түрі, стиліне келгенде бірқатар теориялық мәселе көтереді, күйдің түрлері (формасы) туралы елдің бәрі біле бермейтін «құпияны» ашады.

Ғалым соңғы уақытқа дейін кейбір музыка зерттеушілердің еңбектерінде орын алған: «Қазақ күйлерінен параллельдік кварта-квинталық дыбыстардың үнін ғана естуге болады» - деген ағат пікірлерді жоққа шығарып, Батыс Қазақстан күйлерінде тіпті полифонияның элементі бар екенін дәлелдеп береді.

Көптеген ізденуші қазақ музыка аспаптарын өркениетті түрде танып-білудің бастауын сонау әл-Фарабиден таратады: «Әл-Фарабидің аспаптануға, аспаптық музыкаға қосқан үлесі барша ғалам еліне ортақ, қазіргі аспаптанушылар назарынан да тыс қалмауға тиіс баға жетпес мұра», - деп көрсеткен [7].

Қазақ тіл білімінде алғаш бұл мәселеге қатысты Қ.Жұбанов, Н.Т.Сауранбаев, С.К.Кеңесбаев, Ә.Т.Қайдар және т.б. ғалымдардың зерттеу жүргізгені белгілі [5, 11-б.].

Мен өз баяндамамда осы аталғандардың ішінен тек қана Ахмет

Жұбановтың ағасы, қазақ филологиясының тұңғыш профессоры, қазақ тіл білімінің теориялық негізін қалаушы Құдайберген Жұбановтың 1936 жылы жарық көрген «Қазақ музыкасында күй жанрының пайда болуы жөнінен» атты мақаласындағы кейбір мәліметтерге тоқталмақпын [2].

Ғалым Құдайберген Жұбанов аталған мақалада белгілі бір халықтың рухани мәдениетінің тарихын зерттеуде тілдік материалды пайдалануға болатындығын дәлелдейді. ***Күй, домбыра, қобыз, жыр, толғау, бақсы, сыбызғы, ән, би*** сөздерінің этимологиясын ашу арқылы қазақтың музыка мәдениетінің өте әрідегі тарихына көз жүгіртеді.

Белгілі бір әлеуметтік, саяси-экономикалық, мәдени құбылыстардың тарихына үңілуде, тіл материалдарына қоса сүйенуге болатындығына зерттеушілер көп жағдайда мән бере бермейтіні белгілі. Құдайберген және оның інісі Ахмет Жұбановтар күй мен әннің әуелде бір болғанын, бара-бара музыка мәдениеті жетіле келе, бірінен-бірі ажырап, вокалдық музыка мен инструменттік музыка болып негізделіп шыққанын айтады. Қазақтағы ***күй*** дегеннің халық музыкасының жетілген ірі тарауы екенін дәлелдейді. Қазақтағы музыка терминдері *(домбыра, қобыз т.б.)* кейбір күншығыс тіл мамандары ойлағанындай, бергі замандарда басқа халықтар тілінен ауысқан емес, өте әріде пайда болған төлтума сөздері (түркі сөздері) екенін дәлелдейді. ***«Күй»*** сөзі басқа кейбір түркі тілдерінде ***«көк»*** тұлғасында қолданылатынын айта келіп, XI ғасырдағы Махмуд Қашқаридің сөздігінде ***«Ол ән салып тұр»*** деген сөйлем ***«Ол көкләнер»*** түрінде берілгенін келтіреді.

Түркі-монғол тілдерінде бар ***«ыр»*** мен ***«дола»*** музыка терминдері қазақша ***«жыр, толғау»*** тұлғасында қалыптасқанын да, ***«ән»*** деген сөз иранның ***«оһәник»*** деген сөзінен енген термин екенін де, ***«қобыз»*** сөзі сол инструмент жасалатын материал ***«қамыс»***-пен қатысты екенін де және т.т. Құдайберген Жұбановтың осы мақаласында сөз болады.

Академик Рабиға Сыздықтың сөзімен айтсақ [4, 38-39-бб.]: «Құдайберген Жұбановтың «Қазақ музыкасында күй жанрының пайда болуы жөнінен» атты шағын зерттеуінің тағы бір әсерлі, құнды жағы – осында автор сөз еткен объектісін жақсы білетіндігін, ол жөнінде көп әдебиет оқып, сан алуан пікірлермен таныс екендігін байқатады, зерттеуші бір ретте XIV ғасырда өткен иран музыканты Әбдіқадір Мұрағидің «Зубдәтел әдуар» деген кітабын талдап кетсе, екінші бір жерлерінде түрік музыканты Рауф Ектаның, европа зерттеушісі Кизеветтердің, түрік музыкасын зерттеген проф. Беляев, Успенскийлердің, өзбек музыкасын зерттеген проф. Питраттың, Стамбул университетінің профессоры Гюперли-заденің, «Шығыс музыка гаммалары» деген кітаптың авторы проф. Оганезашвилидің пікірлерін келтіріп, олардың кейбірімен айтысып, кейбірін қоштап өтеді.

Аталған мақаланың тіл ғылымы үшін мән-маңызы қандай болса, музыка мамандары үшін де пайдасы мол деп ойлаймыз, өйткені мұнда музыка жанрының пайда болу, даму, бір-бірінен ажырасып, бөлектену кезеңдері сөз болады».

Ғалым Құдайберген және Ахмет Жұбановтар өз зерттеулерінде әр халықтың мәдениеті, өнері сол этностың танымымен, тілімен, тарихымен тұтастықта дамитынына ерекше зер салады. Тек әуенге негізделетін күйдің де астарында ортаның пәлсапасы, мінез-құлқы көрініс табады деген ғалым ойларының желісі терең логикалық қорытындылармен дәйектелген, яғни: «Халықтың музыкамен байланысты салт-санасын, идеологиясын зерттеген уақытта алдымен ескеріп өтетін нәрсеміз – сол музыканың өзінің шыққан ортасы» болу керек деген тұжырым [2, 280].

Осы айтылғандарға байланысты ізденуші-ғалым Э.Н.Оразалиева «Қ.Жұбанов және шығыс елдерінің танымдық пәлсапасы» атты мақаласында мынандай пайымдаулар жасайды: «Олай болса, музыка этникалық танымның құралы ғана емес, ол сол халықтың психологиясын, бар болмыс-бітімін әлемге таныстырушы символдық құрал болып табылары да сөзсіз», - дейді [2, 280].

Өнер атаулының магиялық құдіретін дәл жеткізу мақсатында ғалым Құдайберген Жұбанов күй мен көк сөздерінің өзара мағыналас табиғатын ұштастырып түсіндіреді, сол арқылы «аспан, көк» ұғымдарының асқақтығын, оған мифологиялық қабылдаудың тигізер әсерін, шығыс халықтарының философиялық ойлауын сұрыптауға тырысқан. Ғалымның сөзімен айтсақ: «...Музыка тіл сияқты әр ұлттың, әр халықтың тарихи мәдениет жөніндегі өзгешеліктеріне қарай өзгеше болады. Сондықтан әр ұлттың, әр халықтың өз музыкасы өзіне жақынырақ, ұғымдырақ болады» [2, 38].

«Бейнемен сөйлейтін» музыка тілінің этнотанымдық ерекшелігіне терең мән берген Құдайберген Жұбановтың аталмыш мақаласы, шын мәнінде, қазақ халқының бар болмыс-табиғатын, этнопсихологиялық әлемін тереңнен пайымдай түсіне білген, даралықтың куәсі.

«Домбыра» атауы, бір қарағанда, қалың жұртқа құпия-сыры мол, томаға-тұйық дүние болып көрінсе де, енді бір қарағанда, барша бітім-болмысы, табиғаты, үлгісі мен өрнегі, үні мен дауысы, дыбысталуы мен сөз құрылымы түгелдей ғасырлар бойы далада өсіп, өмір кешкен қазақ этносының менталитетіне түгелдей сай, сәйкес, үндес келіп жатқан төл құбылыс болып көрінеді.

Қалай жасалып, қандай мағыналық уәжге сүйеніп қалыптасса да, домбыра аспабының атауы қазақ халқының төл туындысы. Мұны негізгі үш ұстанымға сүйеніп дәлелдеуге болады:

1) этникалық-мәдени орта – ол қазақ қауымы;

2) этимологиялық негізі – ол қазақ тілі;

3) аталу уәжі – ол домбыра аспабының дыбыстық ерекшелігі.

Міне, осы үш тұғырға арқа сүйеп қарастырғанда ғана шындыққа жетіп, «домбыра» атауының тілдік табиғатын дәл ашып беруге болады [5, 12-б].

Қазақтың халық аспаптар оркестрін ұйымдастыруда ұлттық аспабымыздың дыбыс күшін ұлғайту үшін көп ізденген ғалым Ахмет Жұбанов халық аспаптарының тарихын зерттеп, қазақ домбырасының түп төркіні шығыс халықтарының музыкалық аспабында жатқанын дәлелдей түсті. Оның бойындағы музыкаға деген ерекше табиғи сезімталдық, әр әнші-

күйшінің ерекшелігін, орындаушылық өнеріндегі өзіне тән шеберлігін, айла-тәсілін қағыс жібермей байқайтын көрегендік, халық музыкасына, халық композиторларының өмірі мен творчествосына байланысты ел аузындағы аңыз әңгімені, тарихи деректерді көкірегіне құйып алатын ұғымталдық, міне, осы қасиеттің бәрі Ахмет Жұбановты халық музыкасының тарихын, оның көркемдік қасиетін, ерекшелігін терең білетін шежірелік, терең ойлы ғалымдық дәрежесіне дейін жеткізді. Халық музыкасының құпия сырын, сұлулығын терең түсініп, батыс және орыс классикалық музыкасынан сусындаған, қазақтың тұңғыш оркестрінің үлкен творчестволық көркем коллектив болып қалыптасуына тікелей әсер еткен ірі ғалым және композитор Ахмет Жұбановтың аты бүгінгі таңда 100 жасқа толған мерей тойын туған топырағында тойлап отырған тек Ақтөбе жұртшылығының ғана емес, бүкіл егеменді еліміздің есінде мәңгі сақталатынына сеніміміз мол демекпін.

Әдебиет

1. Жубанов А.К. Казахский музыкальный инструмент – домбра. К VII международному конгрессу антропологических и этнографических наук АН СССР. Институт им. Н.Н.Миклухо-Маклая. М., 1964, 6-7-бб.

2. Жұбанов Қ.Қ. Қазақ тілі жөніндегі зерттеулер. – Алматы, 1999 ж.

3. Рсалдин Ж. Ахмет Жұбанов. – Алматы: Ғылым, 1966. –110 б.

4. Сыздықова Р. Ғалым-азамат (Профессор Құдайберген Жұбанов туралы). – Алматы: «Арыс» баспасы, 2004.

5. Жаубасова Т.Б. Қазақ ұлттық музыка аспаптары атауларының этнолингвистикалық сипаты. Автореф. ф.ғ.к. – Алматы: Арыс, 2006, 26 б.

6. Оразалиева Э.Н. Қ.Жұбанов және шығыс елдерінің танымдық пәлсапасы // Құдайберген Жұбанов тағылымы: Республикалық ғылыми-практикалық конференция материалдарына арналған мақалалар жинағы. – Алматы, 2004. – 49-55-бб.

7. Бейсенбекұлы О. Сазды аспаптар сыры. – Алматы: Ана тілі. 1994. – 128 б.

ЛЕКСИКОГРАФИЯ САЛАСЫНДАҒЫ АВТОМАТТАНДЫРУ ТӘСІЛІ

ҚР БҒМ А.Байтұрсынұлы атындағы Тіл білімі институты.
«Қазақ тіл біліміндегі функционалдық бағыт» атты
Республикалық ғылыми-теориялық конференция
материалдары. Алматы, 2006, 158-164-бб.

Лексикографиялық нысанның ерекше түрі ретіндегі **автоматты сөздіктер (АС)** мен **автоматы терминологиялық банктердің (АТБ)** пайда болуы бірнеше себептермен түсіндіріледі. Біріншіден, ол электронды есептеу машиналарының (компьютерлердің) мүмкіндіктерінің артуы болса,

екіншіден, лексикографтар мен аудармашылардың еңбек өнімділігін арттыру жолдарын іздестірудің қажеттігінен туындайды [1, 318-333].

Лексикография саласы адамның қызметтері ішіндегі ең бір көп еңбек сіңіруді керек ететін түріне жататыны мәлім. Сөздіктердің кейбір түрін түзу жұмысын сапалы аяқтап шығу үшін оған бірнеше онжылдықтардың өтуі мүмкін. Сондықтан лингвистикалық жұмыстарды автоматтандырудың алғашқы жетістіктерінің лексикография саласына қатысты болуы кездейсоқ жағдай емес. Неше түрлі сөзтізбелерді, үлкен көлемді сөзнұсқағыштарды, жиілік сөздіктер мен кері-әліпбилі сөздіктерді мақсатқа сай өмірге келтіру үшін оларды компьютерлердің көмегімен түзудің қажеттігі қазіргі кезде еш дау туғызбайды.

АС-тардың негізгі міндеті ретінде адамның мәтінмен әрекет етуінде оның еңбек өнімділігін арттыруды жатқызуға болады. Мысалы, лексикографиялық түсініктеме беру мақсатындағы әртүрлі тіл бірліктерін жинақтау, сақтау және өңдеу жұмыстарын **АС** арқылы іске асырған жөн. Яғни сөздіктерді олардың функционалдық критериясы негізінде қарастыратын болсақ, **АС**-ты айырықша сөздік деуге болады. Оның міндеті – мәтінді автоматты өңдеудің (**МАӨ**) практикалық жүйесінде қолданыс табу.

Кез келген **АС**-тың міндетті түрдегі бөлігі ретінде сөзтізбе алынады. Мұндай сөзтізбе **семантикалық баяндау жүйесіне** (**СБЖ**) немесе **білім базасына (ББ)** енудің тетігі болумен бірге тілдік базаның бірден-бір құрастырушысы да бола алады. Мысалы, ақпараттық қызмет ету жүйелерінде түзілетін сөздіктер өз құрамында тиісті білім базасына (тезаурус, фрейм жүйелеріне) ену тетігі рөлін атқаратын дискрипторлық сөзтізбелерді сақтай алады. Ал **машиналық аударма (МА)** жүйелеріндегі сөзтізбелер мәтінге лексикалық және синтаксистік талдауды іске асыру үшін қажетті деген морфологиялық, синтаксистік және семантикалық сипаттамаларға ие **лексикалық бірліктерді (ЛБ)** өз құрамында қорландырады. Осымен бірге мұндай жүйелер әртүрлі мағыналарды, синтаксистік және семантикалық ережелерді сипаттайтын **АС**-тарды де өз құрамына енгізуілері қажет. Мұндай **АС**-тар **МА** жүйесінің **білім базасын** құрайды.

Кез келген эксперименттік жүйе жұмысы да өз құрамында білім базасына жол салатын сөзтізбе болуын қалайды.

Сөйлеу тіліндегі ақпаратты өз жадына ендіруге бейімделген жүйелердегі сөзтізбе құрамында сөзтұлғалардың жазылуы мен оның транскрипциясы да орын алады.

Түсіндірме типті біртілді **АС** сөзтізбесі сөздік жобасының принциптеріне сүйене отырыш анықталады. **Автоматты аударма сөздіктері (ААС)** негізінен ғылыми-техникалық мәтіндермен әрекет етуге арналған болғандықтан, олардың көпшілігі терминологиялық лексиканы қамтулары тиіс. Ірі **ААС** нәтижелігін арттыру үшін оған тақырыптық сөздікке қатысты барлық терминологиялық лексикалар қамтылуы қажет. Яғни жақын тақырыптық аядағы аса жиі қолданыстағы жалпы лексика мен жалпығылымдық, жалпытехникалық лексикалардың барлығы бірдей есепке алынады. Жалпылексикалық қор, ең алдымен, техника саласындағы

мамандарға қажеттірек деуге болады. Себебі ондай мамандар әдетте өз мамаңдығына қатысты лексиканы жақсы меңгергенімен, шет тіліне қатысты лексиканы қажетті дәрежеде біле бермейді.

АС-тың міндеті ғылыми-техникалық мәтіндердің барлық түрлерімен әрекет жасау болғандықтан, олардың құрамына барлық терминдер түрлері енулері қажет. Мысалы, оларға жататындар: стандарттанған және ұсынылған, сол сияқты ұсынылмаған, тұрақтанған немесе жаңадан қолданысқа енген терминдер.

АС-тың құрылымын анықтау жолында мынадай негізгі үш мәселенің шешімі табылғаны жөн:

1) сөздік мақаланың басқы бірлігінің тұрпатын анықтау;

2) **ТБ**-ны сипаттауға қатысты ақпарат жиынтығын анықтау;

3) **АС** базасындағы сөздік мақаланы құрастырудың ұтымды тәсілін таңдау.

АС үшін оның сөзтізбесінің негізгі бірлігін таңдау маңызды мәселе болып саналады. Жалпы лексиканы қамтитын ірі сөздіктерде сөзтізбенің негізгі бірлігі – **негіз сөз** ретінде алынады. Ал әрбір сөзге қатысты сөз тіркестері, әдетте, мақала ішіндегі мағыналық көрініс ретінде немесе басқы тілдік бірліктердің тіркесімдік сипатына мысал ретінде беріледі.

Автоматты аударма сөздікке қатысты сөзтізбенің негізгі бірлігі дәстүрлі терминологиялық сөздіктердің осындай бірліктеріне жақын сипатта болулары қажет. Сөзтізбедегі лексикалық топтың 80 пайызы сөз тіркестерден тұратыны да ескерілуі қажет. Ең ірі **ААС**, көпшілік жағдайда, дәстүрлі терминологиялық сөздіктерге жүктелетін мәселелерді шешуге бағышталған терминологиялық банк рөлін атқарады. Мұндай **ААС**-тарда дәстүрлі терминологиялық сөздіктерге қарағанда сөз тіркестерінің жоғарғы шегі анық емес, яғни оларда сөз тіркестер саны тұрақсыз деуге болады. Ірі **ААС**-тарда дербес мақала ретінде бөлек сөйлемдер мен мәтін үзінділері де есепке алынады.

Аударма сөздіктердегі сөзтізбені іріктеу мен жинақтаудың аталған түрлері әдетте дәстүрлі лексикографияда қолданыла бермейді. Ал автоматты сөздіктер үшін мұндай тәсіл белгілі жағдайда ұтымды болып келеді. Себебі стандартты мәтіндер бойынша жартылай автоматтанған аударма жасау тұстарында **ААС**-тар өздерін тиімді құрал ретінде көрсете бідді. Мысалы, заң құжаттары мәтіндерінде белгілі бір мәтін бөліктері онша өзгеріссіз қайталанып отыратыны белгілі.

Қазіргі кездегі өндірістік **ААС**-тар аса күшті компьютерлер негізінде құрылады. Мұндай компьютерлер бірнеше жүзмыңдаған көлемді сөз тіркестерінен тұратын тілдік ақпараттармен табысты әрекет ете алады. Сонда да компьютер әрекетінің жылдамдығы оның жадының көлеміне тікелей байланысты екендігін ескере отырып, оны үнемді пайдалануды әр уақытта ұмытпау керек.

Екітілді және көптілді лексикография жағдайыңда сөздік бірлігін таңдау нақты тілдердің сөзжасам және тұлғажасам ерекшеліктерімен анықталады.

Айырғыш, аналитикалық, кейбір жағдайда флективті-аналитикалық тіддер типтерін өңдеуге қатысты тез әрекет ететін жүйелер құруда «**Мәтін –**

АС» өзара әрекеттесу сызбасы өзін-өзі ақтап жүргенін айта кеткен жөн. Мұндай жағдайда сөздіктегі белгілі бір сөзге тіркелген барлық ақпаратпен танысу үшін синтаксистік деңгейге дейінгі морфологиялық талдау алгоритмінің көмегіне сүйенбеуге де болады. Сонымен бірге **АС** бойындағы сөзтұлғаларды олардың мәтіндегі «егіздерімен» (сыңарларымен) теңдестіру аса үлкен жылдамдықпен іске асады. Ал синтетикалық, флективтік және агглютинативтік тілдердің **АС**-тарының бас бірлігі ретінде сөзтұлға алынатын болса, сөздік көлемі кенет ұлғаяды да мәтін бойындағы сөзтұлғаны **АС**-тағы сөзтұлғамен теңдестіру жылдамдығы, керісінше, азаяды. Мысалы, орыс тіліндегі зат есім сөз 12-ге дейін, ал етістік сөз, егер түр корреляциясын есепке алатын болсақ, 188 рет тұлға өзгерісіне ұшырайды екен. Түркі тілдерінде есім түріндегі тілдік бірлік саны, теориялық жағынан алғанда, тұлға өзгерісіне шексіздікке дейін ұшырауы мүмкін.

Осы айтылғандарға байланысты мұндай парадигмаларды компрессиялау (қысу) қажеттігі туындайды.

Бұл мәселенің шешімі екі түрлі тәсілмен іске асуы мүмкін:

1) лингвистикалық (алгоритмдік);

2) техникалық.

Линвистикалық тәсілде негізгі назар сөздіктің бірлігін қалыптастыруға бөлінеді. Яғни оның эталонын, үлгісін және ең ыңғайлы баламасын табуға әрекет етіледі. Синтетикалық, флективтік және агглютинативтік тілдер үшін неше түрлі компрессиялау тәсілдері қолданылып жүр. Олардың ішінен ең нәтижелі деген әдіс ретінде **АС**-ты негіз сөздердің сөздігі ретінде қалыптастыру болып табылады. Мұндай **АС**-тарда, әдетте, тілдік бірліктің екі түрі жайлы сөз етуге болады: **негіз сөз** және **ерекше тұлға**, яғни негіздің алмасуы арқылы немесе супплетивті (әртүрлі негізден жасалған бір сөздің тұлғалары) тәсіл арқылы жасалған сөзтұлғалар. **Ерекше тұлға** бірліктер **АС**-тың құрамына негіз сөздермен бірдей дәрежеде енгізіледі.

Сөздік мақаланың құрамы әртүрлі **ААС**-тарда өзгеше және олардың зоналық айырмашылығы 10-нан 99-ға дейін құбылып отырады. Аталған зоналарда әртипті лингвистикалық және экстралингвистикалық ақпараттар тіркеледі.

Әр **ААС** өзінің сөздік мақаласы ішінде белгілі реттілікпен мынадай мәліметтерді қамтиды: басқы сөз, оның бір немесе бірнеше шеттілдік баламалары, баскы сөздің тақырыптық қатыстылығы және оның шеттілдік баламалары, грамматикалық ақпарат, формалданбаған түсініктемелер немесе стандарттанған дефинициялар, мәнмәтіндер, лексикалық тіркесім, басқы сөздің семантикалық иерархиялық байласымы, сапалық коды (сенімдік индексі), қысқартулар, стильдік мәнері (бояюы), лингвогеографиялық шектелуі, тірек сөздер (егер басқы лексикалық бірлік сөз тіркес болса), қайнар көздер, сөздік мақаланың авторы, датасы, әртүрлі қызметтік ақпараттар.

Сөздік мақаланың құрылымындағы семантикалық, синтаксистік және морфологиялық қатынастар арқылы мақаланың өзіне тән жазылу ерекшелігі анықталады.

Деректер базасында (**ДБ**) орын алатын әрбір лексикалық бірлік (**ЛБ**) морфологиялық, синтаксистік, семантикалық және функционалдық деңгейлерде сипаттамалар жиынтығы ретінде көрініс табады. Синтаксистік және морфологиялық сипаттамалар мәтін бойында кездесетін сөзтұлғаның мағынасын анықтау үшін және оның сөйлемдегі синтаксистік қызметін белгілеу мақсатында қажет. Семантикалық сипаттамалар **ЛБ**-ның мағыналық жағын баян етеді де, ал функционалдық сипаттамалар бірліктің барлық құжат құрылымындағы потенциалды рөлін анықтайды.

Қазіргі кездегі мәтінді автоматты өңдеу (**МАӨ**) жүйелерінде **ЛБ**-ның жоғарыдағы сипаттамаларына тиісті орын беру мақсатында сөздік мақалалардың (**СМ**) төмендегідей үш түрі қолданыс табады.

1. Көлемі алдын ала белгіленген, бірыңғай қатал құрылымды **СМ**. Бұл құрылым **ЛБ** тобын сипаттайтын барлық өлшемдерді мақала ішінде қамтуды алдын ала қарастырады. Мұндай тәсіл **ЛБ**-ларды баяндаудың барлығына бірдей қатысты. Оған **ЛБ**-ның семантикалық және функционалдық өлшемдері де қоса беріледі. Аталған тәсіл бойынша **ЛБ**-ны баяндау тек синтаксистік ақпаратпен ғана шектелуі мүмкін. Бұл жағдайда лингвистикалық ақпараттық базаға (**ЛАБ**) семантикалық ақпарат жайлы мәлімет жеке сөздік ретінде енгізіледі.

Осындай мақала құрылымында ақпараттың барлық типтері мен олардың **байттық** көлем-шегі алдын ала анықталып, белгіленеді. Егер ақпараттың нақты типі анықталмаса, оның орны «бос» күйінде сақталуы қажет. Мұндай қатал құрылымды **СМ**-ға мысал ретінде жалпы көлемі 225 байтты алатын SYSTRAN жүйесіндегі мақаланы атауға болады. Бұл жүйенің **АС**-ы негіз сөздер сөздігі ретінде құрылған. Әрбір **СМ** өз бойында мынадай мәліметтерді сақтайды:

Y – лингвистикалық ақпараттық базаға (**ЛАБ**) енгізу кезіндегі **ЛБ**-ны баяндау тәсілі (жаңадан қосу, түзету, жою); А – базалық машиналық негізді баяндау (описание); В – септік жалғаулар жайлы морфологиялық ақпарат, көптік жалғаулардың түр-тұрпаты және т.б.; С – меңгеруге қатысты мәліметі бар синтаксистік кодтар; Д – семантикалық кодтар; Е – сөз алдында келетін көмекші сөздердің кодтары (барлық сөз таптары үшін бұл орын бос болуы қажет); Ғ – балама ерекшеліктерін баяндауды қамтитын негізгі аудармалар.

СМ-ды құрудың мұндай әдісі ондағы ақпаратты жеңіл толықтыру мен үнемі өзгертіп отыруға барынша мүмкіндік тудырады. Себебі **АС**-та әрбір ақпараттың алдын ала белгіленген өз орны болғанымен, оларда алдын ала ескерілмеген жаңаша типті ақпаратты базаға енгізу **СМ**-ды, сөздікті және базаның өзін түбегейлі өзгертуге әкеп соғады. Сондықтан мәтінді автоматты өңдеу (**МАӨ**) жүйелерінде **ЛБ**-ның барлық сипаттамалары толық қамтылғаны жөн болады. Әрине мұндай талап, әсіресе өндірістік жүйелерде **ЭЕМ**-ның қомақты жадын алуымен бірге **АС**-ты орналастырудың қиындық тұстарына әкеп соғады. Солай бола тұра бұл әдіс, көбінде, ақпаратты екінші рет қайталау мүмкіндігінен сақтандырады.

2. **Оң жақ шегі жылжымалы болып келетін сөздік мақала.** Мұндай мақаланың баяндалуы – «таңба – мағына» принципі бойынша іске

асады. Бұл тәсіл бойынша мақаланы баяндауды қалыптастыру бірізді сызба арқылы жүргізіледі және **ЛБ**-ға қатысты барлық ақпаратты зоналарға үлестіру жолдары қарастырылады. Ондай зоналардың саны нақты мәліметтердің типіне байланысты болып келеді. Ал лингвистикалық бірлікті баяндау үшін, сөзді топтастыру тұрғысынан алғанда, лексика-грамматикалық, морфологиялық, семантика-синтаксистік, тұрпаттық, функционалдық және т.б. көзқарастағы сипаттамаларға сүйенеді. Әрбір зонада **ЛБ**-ның қандай топқа (класқа) қатыстылығына қарай, белгілер саны жекеленіп алынады. Мысалы, зат есімді актуалдаудың функциясын атқаратын детерминативтер класы осы кластың нақты бірлігін баяндайтын мынадай белгілер жиынтығымен сипатталады: сөздің ішкі топтары, өзгеріп түрленуі, орналасу қалпы және т.б.

Сөздік мақаланың «оң жақ шегінің жылжымалы» болуы мақаланы тек **ЭЕМ**-ға (компьютерге) дайындау кезіне ғана сәйкес келеді. Ал компьютердің жадына енгізу тұсында мақаланың баяндалуы, алғашқыдай қатаң құрылымға ауыстырылуы керек.

3. Иерархиялық құрылымдағы сөздік мақала. Бұл тәсіл ақпарат бөліктерін жеке файлдар (массивтер) ретінде бөліктеп шығаруды алдын ала көздейді. Мұндай жағдайда **АС**, ең аз дегенде, екіге бөлінеді: міндетті түрдегі тіркелген ақпараты бар сөздікке ену тізімі және мағыналар сөздігі.

ЛБ-ның мәтінді автоматты өңдеу (**МАӨ**) жүйесінде қолданылу мүмкіндігін ескере отырып, оның барлық сипаттамаларын иерархиялық топтастыру алдын ала жүргізіледі, яғни иерархиялық жоғарғы және төменгі деңгейлері анықталады. **МАӨ**-нің барлық түрлерінде тікелей қодданылатын жоғарғы деңгейдегі сипаттамалар мақала құрылымына тікелей енгізіледі, ал барлық басқа сипаттамалар (әдетте семантикалық) бөлек файл түрінде жекешеленеді. Мұндай жағдайда сөздік мақалада жекешеленген файлдарға арнайы сілтеме беріледі.

Базалық вариантта **СМ** құрылымындағы «кіру» (немесе «ену») сөздігі құрамына жататындар: басқы **ЛБ**-ның өзі (негіз, сөзтұлға немесе сөз тіркесі); ендірілетін тілдің типологиялық ерекшелігіне, грамматикасына байланысты лексика-синтаксистік код; сөздің әртүрлі заттық аяда атқаратын қызметінің ерекшелігін анықтайтын ішкі тіл (подъязык) коды; аударма жайлы ақпарат.

Баяңдаудың жоғарғы деңгейі, яғни сөзтізбе өз құрамында **ЛБ**-ны және сөйлемге алғашқы синтаксистік (синтагмалық) талдау жүргізуге қажетті морфологиялық сипаттамаларды қамтиды. Мысалы, сұрақ-жауап жүйесінде баяндаудың жоғарғы деңгейі – әрбіреуіне синтаксистік және морфологиялық ақпарат тіркелген лингвистикалық бірліктің (**ЛБ**) өзінен ғана тұруы мүмкін. Бұл деңгей анықтағыш деп аталып, іштей екі бөліктен тұрады [1, 324].

Тілдің семантикалық жағын баянды ететін **АС** дербес модуль ретінде бірнеше тәсілдермен құрылады:

а) белгілі ережелер негізінде біріктірілген семантикалық белгілердің жиынтығы түрінде;

ә) кеңейтілген немесе ассоцияланған (байласымды) семантикалық желілер ретінде.

Әдетте, күрделі **автоматты аударма жүйелерінде (ААЖ)** бірнеше тақырыптық аялардың лексикалық топтары қамтылуы керек. Себебі әр тақырыпқа қатысты лексиканың бір-бірімен қиылысу мәселесінің шешімін табу **ААЖ**-да аса маңызды іс болып саналады. Өндірістік **ААЖ**-ды құрастырушылар аталған мәселеге үлкен мән беріп, осы мақсаттың шешімін табуға арнайы арналған лексиканы тақырыптық топтау жүйелерін іске қосуда. **ААЖ**-дағы лексиканы тақырыптық топтауға негіз ретінде ғылыми-техникалық ақпараттың Мемлекеттік автоматтандырылған жүйесі орталығының басқы сөздері (рубрикаторлары) алынған.

Машиналық аударманың (**МА**) автоматты сөздігінің (**АС**) қатаң түрде нақты тақырыптық аяға (**ТА**) бағышталуы лексикалық талдау деңгейінде тіддік бірліктердің (**ТБ**) көпмағыналығын ажыратуға және терминдерді аудару мәселесін стандарттауға мүмкіндік тудырады.

Нақты **ТА**-ға бағышталу **АС**-тың кез келген жүйесі үшін бірден-бір қажет. Ал егер **АС**-ның көпаспектілі жүйесін құру керек болса, оңда **ТБ**-ның нақты **ТА**-ға қатыстылығын ажырату қажеттігінен **СМ**-ның құрылымына арнайы шартты белгі-код енгізіледі. Сол код арқылы лексикалық талдау кезінде тіддік бірліктің аударма баламасы таңдалады.

Дәстүрлі аударма сөздіктеріне қарағанда **АС** динамикалық қасиетке ие, яғни олар сөздіктің сөзтізбесі мен сөздік мақаласының көлемділігіне қарамай, оларға қажетті түзетулерді тез және еш қиыңдықсыз енгізе алады.

АС-ты құрастырушылардың үнемі лексикалық топты таңдау жұмысымен айналысуының арқасында, сөздіктің халық алдында қызмет ету кезеңдерінің бәрінде де өз жалғасын тауып отырады. Қажетсіз деген сөздерді жүйеден шығарып тастап және оны жаңа лексикамен толықтырып, сөздік мақаланың кез келген тұстарын әрі түзетіп, әрі толықтырып отыру мүмкіндігі тұрақты түрде орын алады. Құрамының айнымалығы мен сөздіктің қызметі кезіндегі түрлі түзетулерді қабылдау мүмкіндігі – машиналық сөздіктердің ауытқымайтын (принциптік) және дәстүрлі сөздіктерден (қағаз түріндегі) айырым табатын бірден-бір сипаттамасы.

Соңғы жылдары аса кең орын ала бастаған сөздік – **ЭЕМ** арқылы құрастырылатын конкорданс-сөздіктер. Қазіргі кезде Шекспир, Ронсар, Пушкин және т.б. классиктердің шығармалары бойынша іргелі конкорданс-сөздіктермен **ЭЕМ** арқылы танысуға мүмкіндік туды.

Сөзнұсқағыш сөздік дегеніміз әліпби ретімен реттеліп жазылған және сөздердің қандай мәтіндерде қолданылғынына сілтеме жасалатын сөз немесе сөзтүлғалар сөздігі. Конкорданс-сөздік аталған мәліметтерді өз бойында ұстайды және сонымен бірге әрбір сөздің қолданыс тапқан мәнмәтіндері ол сөздікте қоса беріледі. Түсіндірме сөздіктерге қарағанда конкорданс-сөздіктерде сөз мағынасын ашуға қатысты түсініктеме беру және қажетті деген мысалдармен қамтамасыз ету қарастырылмайды.

ЭЕМ көмегімен түзілген **конкорданс** лексикографты көптеп еңбек сіңіруді қажет ететін әрекеттерден босатады. Мысалы, оған жататындар: сөзтізбені таңдау мен құрастыру, мәнмәтін кәртішкілер қорын жинақтау және т.б. Егер кез келген конкорданс-сөздігіне лексикограф күшімен қажетті деген

толықтырулар жасалынатын болса, ондай сөздік түсіндірме сөздіктердің түр-түрін құрастыруға негіз бола алады.

Конкорданстарды құрастыру жұмыстарын талдай келе, мынадай тұжырым жасауға болады. Қазіргі кезде ең тиімдісі – арнайы және толық түрде автоматтанған сөздік-анықтағыш қызметін атқаратын «сұрақ-жауап» принципінде құрылған конкорданс типтес сөздікті жобалау және оны іске қосу. Осындай жұмыстың негізінде тұтынушының қажетті деген шығармалардың мәтінін **ЭЕМ** жадына бір-ақ рет енгізіп, одан кейін өз қалауынша автоматтанған жүйеге, яғни конкорданскa сан рет «кіруіне» мүмкіндігі бар. **«Деректер банкісі»** (**ДБ**) деп аталатын мұндай жүйелер мәтін ішіндегі ақпараттың тек қана дәлдігін, сенімділігін және дұрыстығын тексеріп қоймай, сонымен бірге тиісті сөздің барлық қодданысы жайлы, әрбір сөзге (сөзтұлғаға) қатысты барлық мәнмәтін түрлері жайлы және шығармалар жиынтығы жайлы толық мәліметтер бере алады.

Мұндай автоматтанған сөздік-анықтағыш қызметі ежелгі жазба ескерткіштерге, жеке авторлардың классикалық әдеби шығармаларына, арнайы терминологиялық жүйе көрінісіне, әр тілде жазылған мәтіндерге барынша бейімделе алады.

Сөздік-анықтағыш қызметін атқаратын автоматтанған лингвистикалық деректер қорын құрудың актуалдылығы еш күмән туғызбайды және мұндай мүмкіндіктің туар күні де алыс болмауы керек.

Сонымен, сөздік-анықтағыш жүйесін құру болсын немесе конкорданстар жүйесін құру болсын бұлардың бәрі де актуалды жалпыфилиологиялық мәселе екені айқын. Мұндай жүйелерді ойдағыдай іске асыру үшін тілшілердің, әдебиетшілердің, математикалық және қодданбалы лингвистика мамандарының тығыз түрде одақтасып, бірлесе жұмыс атқарулары қажет деп білеміз. Болашақта түсіндірме және аударма сөздіктерді, анықтағыштарды, рубрикаторлар мен классификаторларды құрастыру жұмыстарының және терминологияны қалыпқа келтіру, терминологиялық комиссия жұмысына қызмет ету, ғылыми-техникалық аударма жұмыстарын автоматтандыру мәселелері қазақ, лексикография саласынан да орын алатынына сенімізміз мол. Қазіргі кездегі есептеу техникасының, қолданбалы лингвистика мен ақпараттану салаларының табыстары мен жетістіктері мұндай автоматтанған жүйелердің үнемді және тиімді болу мүмкіндіктерінің бірден-бір кепілі деуге болады.

Әдебиеттер:

1. *Беляева Л.Н., Герд А.С, Убин И.И.* Автоматизация в лексикографии. «Прикладное языкознание» учебник. Отв. ред. А.С. Герд. С.-Петербург, Изд.во С.Петербурского университета, 1996. –С.318-333.

2. Лингвистическая концепция терминологического банка данных машинного фонда русского языка: (Проект) / Под ред. А.С.Герда. –М., 1989.

3. *Марчук Ю.Н.* Проблемы машинного перевода. –М., 1983. –232 с.

4. *Марчук Ю.Н.* Машинный перевод как информационная и технологическая реальность. –М., 1988.

5. *Перстнев И.П., Волкова И.Н.* Банки терминологических данных. –М., 1987.
6. *Попов Э.В.* Общение с ЭВМ на естественном языке. –М., 1982. –360 с.

ҚҰДАЙБЕРГЕН ЖҰБАНОВ ЗЕРТТЕУЛЕРІНДЕГІ МОДЕЛЬДЕУ ӘДІСІНІҢ КӨРІНІСІ

ҚР ҒБМ ҒК А.Байтұрсынұлы атындағы Тіл білімі институтының «Тілтаным» журналы. 2006. № 1. 3-9-бб.

Профессор Қ.Жұбановтың филологияға қосқан теориялық және практикалық ғылыми үлестері қазақ тіл білімінің барлық салаларына әсер етті деуге болады. Қазақ тілінің зерттелу болашағын өзіне тән көрегендікпен, табиғи дарындылықпен болжам жасап, олардың кейбір тұстарының ғалым зерттеулерінде орын алғаны жайлы сөз етпекпіз.

Ғалымның осындай зерттеу тұстары тілдің қолданбалы саласының жаңа бағыттарынан да байқалады. Дәлірек айтсақ, тіл зерттеу аясындағы XX ғ. 60-жылдарында ғана пайда болған «Статистикалық лингвистика», «Лингвистикадағы модельдеу» бағыттарының кейбір бастамаларын Құдайберген Жұбановтың сол кездегі зерттеулерінің өзінен-ақ молынан кездестіруге болады.

Қ.Жұбанов өткен ғасырдың 30-жылдарының өзінде-ақ өз зерттеулерін заманға сай жүргізуге ынталанғанын «Қазақ тілінің жоғарғы курсы» атты жазған еңбегіндегі мына тұжырымнан байқауға болады [1, 79-б.]: «XIX ғасырда жаратылыс ғылымы ілгеріледі; оларды зерттеуге түрлі метод қолданылды. Объектив әдісімен іс істеу тілшілерге де жұқты. Бұрын тіл логика негізді әдіспен зерттелетін еді (яғни, сағызша созыла беретін белгілі өлшеуге келмей, көбіне субъектив жорамалға сүйенетін)», – дейді.

Ғалымның сол кездің өзінде-ақ тіл білімінде сандық деректерді пайдаланудың қажеттілігіне де ерекше ден қойғаны байқалады. Мәселен, ғалымның 1936 жылы жарық көрген «Қазақ тілінің грамматикасы» атты еңбегінде сөз мүшелерінің орны, дыбыстардың саны мен сапасы, түрлері туралы жазған тұжырымы осыған дәлел: «Дыбыстың саны бар да, сапасы бар. Дыбыстың сапасы бірдей бола тұрып, саны әлденешеу бола береді.

… Бір тілде жұмсалатын дыбыстардың жалпы санының ұшы-қиыры жоқ, бірақ оның бәрінің бірдей сапасы әр түрлі бола бермейді. Сапасы әр түрлі болатын дыбыстар белгілі мөлшерде ғана болады да, көп болмайды… Қазақ тілінде де сапасы әр түрлі болатын дыбыстардың арнаулы саны бар…» [1, 183-б.]. Тілдік құбылыстарда сан мен сапаны ажырату қажеттігін айта келе, ғалым тұрпат (форма) пен мағынаны да ажырата білудің маңыздылығын тиянақтап айта білген. Мысалы, «Қазақ тілінің жоғарғы курсында» былай дейді [1, 79-б.]: «…Бірақ (грамматикалық құрылымдардың) сыртқы белгілеріне сүйенуден формализм туды. Сөйтіп морфология жеке сөздердің формасын, синтаксис сөз тіркестерінің формасын тексеретін

ғылым деп танылды. … Синтаксистің өзі де … тіркескен сөздердің тіркесу формасын қарайтын ғылым болды» – дей келе, бұл әдісті толық қолдамайтынын білдіреді. Осындай күмәнін дәлелдей түсу мақсатымен және ол бағытты толығынан терістемейтінін білдіру үшін ғалым «Қазақ тілі грамматикасының кейбір мәселелері» атты еңбегінде тағы да былай дейді [1, 34-б.]: «…Рас, біз формашыларға қарсымыз: тіл құбылыстарын жалаң түр жағын ғана зерттеумен біліп болмайды. Бұл – сөзсіз сыңаржақтық. Сонда да жоғынан бары игі. Түр – форма деген тілде жоқ белгі емес, бар нәрсе. Бар қасиетін көрсетушілік қашанда пайда болмаса, зиян келтірмейді. Бірақ жеткіліксіз, өйткені түр (форма) тілдің бар тарауы емес, бір тарауы».

Шындығында да, ғалым зерттеулерінде тілдің формалық сипатына және объективтілікке сүйенетін жақтары баршылық.

Енді, қазіргі кезде қолданбалы тіл білімінің дәстүрлі бағытымен қатар дамып келе жатқан және заман сұранысына сай пайда болған оның жаңаша бағытындағы «тілді модельдеу» әдісінің кейбір теориялық тұстарымен салыстырмалы пайымдаулар жасайық [2; 3; 4].

Модельдеу әдісінің қажеттігі, ең алдымен, зерттеу нысанымызды тікелей байқауға (яғни қолмен ұстап, көзбен көруге болмайтын) жағдайлардан туындайды. Сондықтан осындай модельдеу әдісі, ол өзінше бір бөлек әдіс ретінде, көптеген ғылымдар саласында, соның ішінде қазіргі кезде қазақ тілін зерттеуде де қолданыс тауып жүргені белгілі.

Ғылыми әдебиеттерде «модель» терминіне мынандай анықтама беріледі. Модель дегеніміз – зерттеу нысанының қасиеттерін бойына сақтай алатын және түпнұсқаның (нысанның) орнына жүретін, танымдық процесінің қажеттігінен туындайтын материалдық не абстракты нысан (объект) делінген [2].

Модель әр уақытта өзінің түпнұсқа түріндегі шындығынан да артық күйде, яғни «дәріптелген» сипатта қарастырылады. Осының негізінде, қай ғылым саласының болсын, формалды модельдері математикалық жүйе деп қарастырылады. Сондықтан, басқаша айтқанда, формалдылықтың түсінігі – математикалық дәлдік немесе бірізділік деуге болады. Бұл қасиеттер тіл табиғатына да тән құбылыс.

Осымен қатар, лингвистикалық модель эксплонаторлық қасиетке де ие болу керек деп ұйғарылады, яғни оның түсініктеме беру мүмкіндігі болуы қажет. Аталған қасиетке ие болуы үшін, ол, ең алдымен, осыған дейінгі тілдік теория негізінде түсінігі күңгірт не түсініксіз болып жүрген тілдік құбылыстарды бірізді түсіндіре алатын, анықтама бере алатын мүмкіндігі болуы керек. Сол сияқты, зерттеу нысанының бұрын-соңды ескерілмей жүрген жаңа мүмкіндіктеріне де ғылыми болжам жасай алатын қабілеті болуы қажет деп пайымдалады.

«Лингвистикалық модельдеуде» сөйлеу қызметі моделінің пішінін екі түрде қарастырады. Оның біріншісі – семантикасыз немесе тек грамматиканы игеруге ғана «еліктеуші» таза синтаксистік модель. Екіншісі – мағыналы сөйлемдерді түсіндіре білумен қатар, оларды дұрыс құра да білу мүмкіндігіне еліктейтін – семантикалық модельдер [2].

Мұндай модельдерді құру жазба не сөйлеу мәтіндерін автоматты түрде өңдеуде аса маңызды. Мәселен, «машиналық аударма» жүйелерінің барлық дамыған түрлерінде «автоматтандырылған синтаксистік талдау» жүргізу сөйлемнің семантикалық түсінігін алу үшін және екінші тілдік жүйедегі сөйлемдер құрылымын «аударма грамматикасына» ауыстыру үшін аса қажет деп саналады. Сол сияқты, «ақпараттық іздеу» жүйесінде де толық синтаксистік талдау сөйлем құрамындағы «есімді сөз тіркестері» негізінде танылады. Ал «жасанды интеллект» жасау мәселесіндегі мәтінді тану жүйесіне қатысты синтаксистік талдау, өзінше бір бөлек семантика-синтаксистік топтаманы құрайды деуге болады.

Қазіргі кезде, қай тілдің болмасын, оған тән сөйлем құрылымының дәл және формалды берілісін білу аса қажет болуда. Ол үшін реттелген сызықтық тізбек құрайтын кесінділер жиыны – **«құрастырушылар жүйесін»** немесе **«құрастырушылар моделін»** құра білу керек. Аталған модельде синтаксистік бірлік ретінде сөзтұлғалар, сөз тіркестер және олардың құрылымдық қатынастары қарастырылады.

Тілдік жүйе жиынындағы қатынастар **«граф»** деп аталатын сызба түрінде көрініс табады да және оның төбешігіндегі ең кіші бірлік ретінде не негіз сөз, не сөзтұлға, не сөз тіркес алынады. Сөйлемдердің графтық сызбасын салудың алдында сөйлемдегі әрбір сөзқолданыстар өз реті бойынша нөмірленеді. Әрі қарай, сызбалар түрінде граф төбешігіндегі құрастырушы сөзқолданыстар мағыналық жағынан ішкі сөз тіркестерге бөлектеліп, олар бір-бірімен ішкі және сыртқы тәуелділіктеріне қарай, түзу сызықтық кесінділер арқылы өзара қосылады. Осындай сызба пішіні әдебиеттерде **«құрастырушылар моделі»** немесе **«синтаксистік модель»** деген терминдермен аталып жүр [3].

Сөйлем құрастырушыларына талдау жүргізудің өзіндік қиыншылығына қарамай, **«құрастырушылар модельдері»** сөйлем құрылымын грамматикалық және семантикалық тұрғыда да дұрыс сипаттайды деп саналады. Мұндай дұрыстық сипат, көбінде, мағына ажырату кезінде, мысалы, құрастырушы сөзтұлғалар – **«омонимдес сөздер»** немесе сөйлемдер – **«омонимдес сөйлемдер»** болған жағдайларда анығырақ байқалатыны дәлелденеді.

Сөйлемдер ішіндегі сөзтұлғалар арасындағы **қабысу, жанасу, қиысу, меңгеру** сияқты тәуелділіктер жиыны сол сөйлемнің синтаксистік құрылымы жайлы көптеген мәліметтер бере алады. Ал тәжірибе жүзіндегі ізденістерде, мұндай тәуелділіктердің жиыны граф пішініндегі сызба түрінде көрініс табады.

Профессор Құдайберген Жұбановтың сөйлем құрылымына қатысты зерттеулерінде сөйлем ішіндегі сөзтұлғалар мен сөз тіркестер «граф» немесе «құрастырушылар моделі» деген терминдермен аталмаса да, сөйлемге енетін сөздердің әрбіреуін тиісті нөмірлермен таңбалау және олардың тәуелділіктерін сызба түрінде кесінділер мен төртбұрыштар арқылы беру молынан кездеседі.

Мәселен, ғалымның «О формах сочетания слов в казахском языке» атты мақаласында 15 сөзқолданыстан тұратын мындай сөйлемді қарастырады [1, 476-б.]: *«Сен(1) көрген(2) қымбат(3) алпауыттан(4) да(5) Қызыл(6) құрт(7) Ертай(8) ұстаның(9) Темірбайдан(10) алған(11) төрт(12) бесті(13) аты(14) жақсы(15)».* Осы сөйлемдегі сөздердің орын ауысуынан сөйлем мағынасының өзгеру заңдылығын көрсету мақсатымен, әрбір сөзді нөмірлеп алып, бес түрлі айтылым ретінде бөлек-бөлек қарастырып, пайымдаулар жасайды. Сөйлемді орыс тіліне былайша аударады: *«Чем(4) дорогой(3) помещик(4) виденный(2) тобою(1), лучше(15) и(5) гнедой(12) пятилетний(13) мерин(14), полученный(11) кузнецом(9) Ертаем(8) кызыл-куртовцем(6) от(7) Темирбая(10)».*

Әрі қарай осы сөйлемдегі сөздерді түпнұсқадағы баламаларымен салыстыра отыра, олардың мағыналары мен тұлға өзгерістерін талдайды. Міне, мұндай әрекеттер қазіргі кезде арнайы әдіс ретінде сөйлем синтаксисін модельдеуге қатысты зерттеулерде жиі қолданыс тауып жүр.

Қ.Жұбановтың зерттеулерінде, яғни «Сөйлем мүшелері» [1, 153-157-бб.], **«**Жетек сөз бен жетекші сөз» [1, 160-161-бб.], «Сөйлем мүшелерінің орын-орны» [1, 162-б.] деген тақырыпшаларда «О формах сочетания слов в казахском языке» [1, 476-481-бб.], «О построении речи на казахском языке» [1, 482-495-бб.] және «Этюды к переводу трех стилей» [1, 531-534-бб.] атты мақалаларында көрініс тапқан сызбалар үлгісі 12 түрлі сурет арқылы берілген. Бұл келтірілген сызбаларды қолданбалы лингвистика саласының терминімен айтсақ, белгілі мағынада, **«Қазақ сөйлемдерінің синтаксистік модельдерінің үлгілері»** деуге де болады.

Енді сөйлем ішіндегі тәуелділіктердің кесінді сызба түрінде көрініс табуын Қ.Жұбановтың аталған еңбегінде мынадай сөз орамдарын байқауға болады:

1) кітаптың 156-бетінде: *«Айқындауышты өзінің айқындап тұрған сөзімен сызық арқылы байланыстырып көрсетсек мынадай болады»* деп екі түрлі сызбаны ұсынады;

2) кітаптың 160-бетінде: *«Біздер мектепке келдік»* деген сөйлемді мысал ретінде алып, былай пайымдаған: «… бұл үш сөз матастырылмаған. *Біздер* деген мен *мектепке* деген екеуі жиналып *келдік* деген бір сөзге арқандалған. Мұның арқандалу ретін сызықпен көрсетсек мынадай болады» деп және *«Енді қай сөз, қай сөзді жетектеп тұрғанын да сызбамен таңбаласақ, мынадай болады»* деп тиісті сызбаларын ұсынады;

3) кітаптың 162-бетіндегі «Сөйлем мүшелерінің орын-орны» деген тақырыпшада бастауыш, баяндауыш, анықтауыш және толықтауыш болатын сөйлем мүшелерінің орындарын көрсету үшін М.Әуезов шығармасынан алынған мына сөйлемнің *«Бұл – мал күткен, малға ықылас берген бейілдің көрмесі»* әр сөзін нөмірлеп алып, былай дейді: *«Мұндағы сөздердің қалайша қиындасып тұрғанын сызықпен көрсетсек, мынандай болады»* деп осыған сәйкес салынған сызбаға сілтеме жасайды.

4) кітаптың 482 бетіндегі «О построении речи на казахском языке» атты мақаласында 16 сөзқолданыстан тұратын сөйлемді қарастырып,

олардың формалды белгілері арқылы және бір сөзге қатысты параллель тіркестердің орын тәртібінің қатал сақталмай, тек логикалық екпінге ғана бағыныштылығы жайлы сызбаны ұсынады. Төменде *1-суретте* орын алған бұл сызбаның сөйлемнің синтаксистік моделінің көрінісі екендігіне осы саланың мамандары күмән туғызбас деген ойдамыз.

5) келесі қарастыратынымыз ғалым кітабының 531-бетіндегі көлемі жағынан шағын (4 бет), ал құндылығы жағынан «аударманың синтаксистік және семантикалық моделі» ғылыми бағыты үшін аса маңызды «Этюды к переводу трех стилей» атты (тақырыбы шартты) атты зерттеу жұмысы. Бұл мақала ҚазССР-ның Орталық архивінен алынып, тұңғыш рет Алматыда «Лексико-семантические вопросы перевода» атты жинақта 1987 жылы басылып шыққан болатын. Мұның құндылығы үш түрлі стиль (көркем әдебиет, ғылыми және публицистикалық) мәтіндеріндегі сөйлемдерді салыстыра-салғастыра қарастыруында.

Әр стильге тән ерекшеліктерге қатысты сөйлемдер ішіндегі сөздердің бір-бірімен байланысын 5 түрлі сызбалар арқылы орыс тіліндегі сөйлемдерді қазақ тіліне аударуға қатысты өзіндік семантика-синтаксистік модель-үлгіні ұсынады. Осы сызбалар ішінен біз тек көркем әдебиет мәтініне қатыстысын ғана, яғни *2-суреттегі* сызбаны оқырман назарына ұсынып отырмыз.

Ғалым Құдайберген Жұбанов аталған сызбаларына сілтеме жасау үшін мынандай айтылымдарды қолданған: *«сызық арқылы байланыстырып көрсетсек», «мұның арқандалу ретін сызықпен көрсетсек», «қай сөз, қай сөзді жетектеп тұрғанын да сызбамен таңбаласақ», «сөздердің қалайша қиындасып тұрғанын сызықпен көрсетсек»* және т.б.

Тілді оның формалды пішіні арқылы сипаттау «алмасу ақпаратының» мазмұндық сырын ашуға толық мүмкіндік туғыза бермейтіндіктен зерттеуге тиісті ақпараттың мағыналық, яғни семантикалық жағын бөлек қарастыру қажеттігі туындайды.

Көптеген зерттеушілердің тұжырымдауынша, сөйлем мағынасы аса күрделі және көпаспектілі құрылым. Сондықтан синтаксистік модельдеу барысында ұстаным етілетін төрт түрлі ғылыми бағыттың үшеуі **«әлем–ойлау–тіл»** үшбұрышы элементтерінің қатыстылығына қарай, ал төртіншісі – **«тіл–адам»** прагматикалық қатынасына негізделеді. Ғылыми әдебиеттерде бұл бағыттар: *онтологоцентристік, концептоцентристік, синтактикоцентристік* және *антропоцентристік бағыт* деген терминдермен аталып жүр.

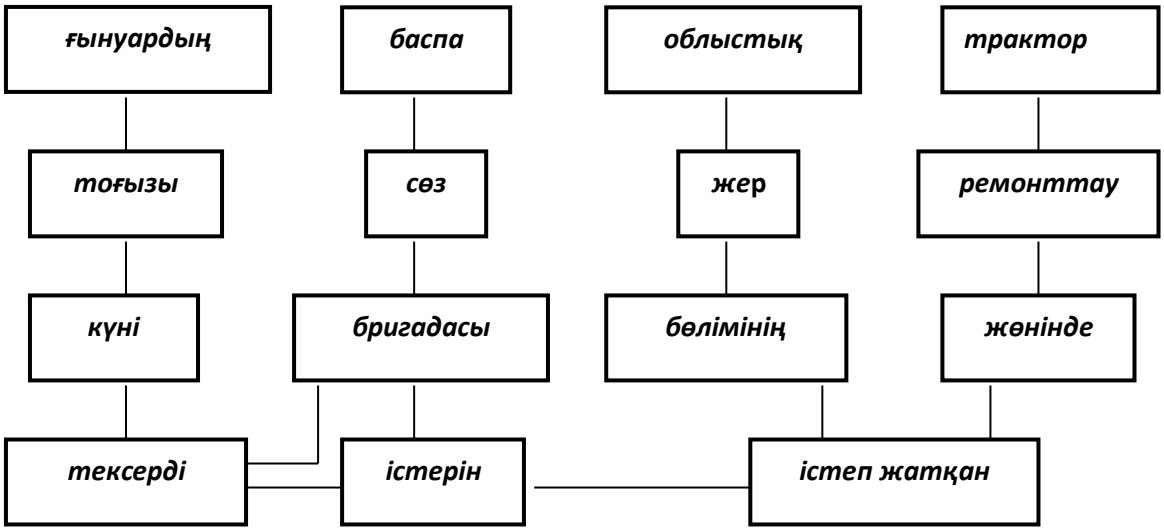
Қазіргі кезде *онтологоцентристік* және *концептоцентристік* бағыттар бойынша коммуниканттар факторын ескермеу және коммуникация жағдайымен санаспау сөйлем семантикасын дұрыс ашуда нәтижелі болмайтындығы да дәлелденді.

Сондықтан бұл жағдайларды ескеру «лингвистикалық прагматика» деген терминмен аталып, айтылым мазмұнына семантикалық компонентпен бірге прагматикалық компоненттің де енуін қажет деп санайды. Оған жататындар: сөйлеуші факторы, тыңдаушы факторы және сөйлеу актісіне

қатысы бар – орын мен уақыт, сөйлеуші мен тыңдаушының әлемдік көзқарасы, жалпы білім қоры және т.б.

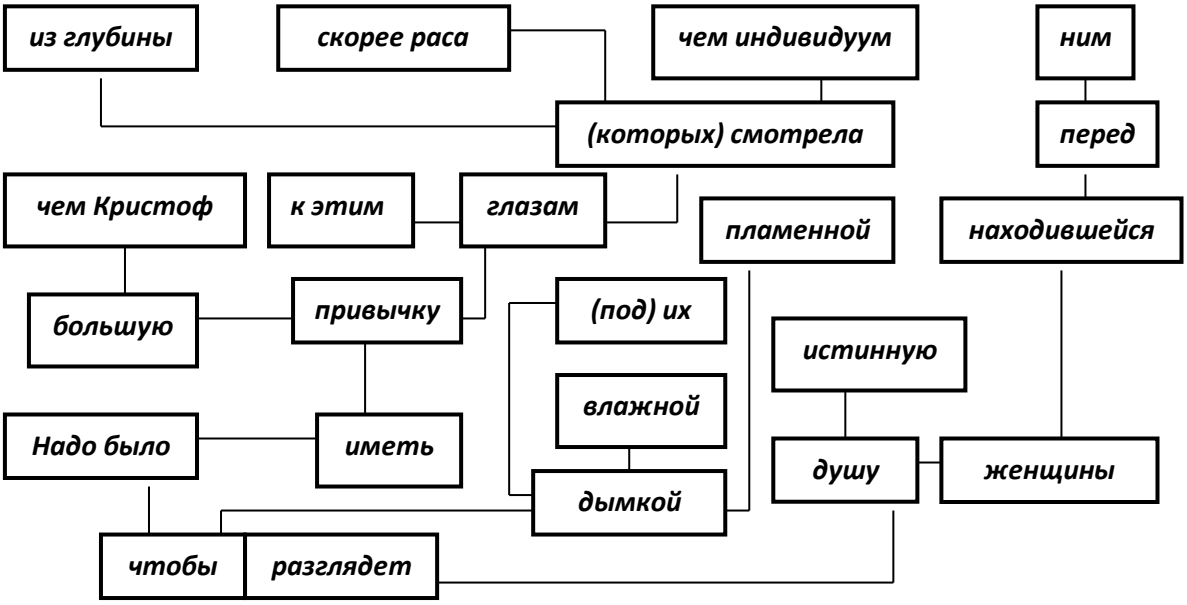
Ғылыми әдебиеттерде сөйлеушіге қатысты айтылым үлгілері – «сөйлеуші моделі», ал қабылдаушыға не тыңдаушыға қатысы тілдік мәселелер – «тыңдаушы моделі» деген терминдермен аталып жүр. Сөйлеуші-коммуниканттың рөлі, ең алдымен, айтылым арқылы белгілі мақсатқа жетуге ниеттену үшін иллокутивті функцияларды пайдалануды қажет ететіні белгілі. Ал иллокутивті функция сөйлеушінің ниетін білдірумен қатар ол тыңдаушымен де санасуы қажет етеді екен.

«Трактор ремонты жөнінде облыстық жер бөлімінің істеп жатқан істерін баспасөз бригадасы ғынуардың тоғызы күні тексерді».



1-сурет

Схема связей слов



2-сурет

I. Текст художественный (Ромен Роллан. Жан Кристоф. Т.4. С. 68).

а) «Надо было иметь большую, чем Кристоф, привычку к этим глазам, из глубины которых смотрела скорее раса, чем индивидуум, чтобы разглядеть под их влажной и пламенной дымкой истинную душу находившейся перед ним женщины».

Перевод: *«Алдында тұрған әйелдің шыны кім екенін көзінің буы мен жалыны аралас түтінінен айырып тану үшін, бұл аржағынан (түбінен) бір адам емес, бүтін ұрулы ел қарап тұрған екі көзге Кристофшыламай әбден үйреніп алу керек еді».*

Міне, осындай (*онтологоцентристік, концептоцентристік, синтактико-центристік, антропоцентристік* және т.б.) толып жатқан басқа да терминдер арқылы аталып жүрген «қолданбалы тіл білімінің» жаңа бағытына қатысты мәселелер Қ.Жұбановтың зерттеулерінде дәл осылайша аталмағанымен, өзіндік мазмұнда молынан көрініс тапқан.

Құдайберген Жұбановтың зерттеулерінде орын алған бұл жағдаяттар қазақ тілінің қолданбалы саласының жаңа бағыты бойынша жүргізілетін зерттеулерде өз жалғасын табатыны мәлім.

Сонымен, қазақ тілінің теориялық және қолданбалы тіл білімінің негізін қалаушы, қазақ филологиясының тұңғыш профессоры Құдайберген Қуанұлы Жұбановтың азғана өмірінің бір-ақ мақсаты болған сияқты. Ол – әлі ашылмаған шындықты ашу немесе соның ашылуына ықпал жасау, сол арқылы қоғамға қызмет ету. Мұндай мақсатқа жетудің әр түрлі жолдары бар екені белгілі, бірақ ғалым Қ.Жұбанов солардың ішінен зерттеудің дәлдігі мен объективтілігін қамтамасыз ететін тәсілді ғана ең дұрысы деп тапқан болу керек. Шындығында да, дәлдік пен объективтілікке ұмтылу – қазіргі ғылымдарға бірден-бір тән қасиет, ал бұның тіл біліміне де қатысы бары сөзсіз.

Әдебиет

1. Қ.Жұбанов. Қазақ тілі жөніндегі зерттеулер. – Алматы: Ғылым, 1999. – 581 б.

2. М.И.Откупщикова. Моделирование языка // Прикладное языкознание: Учебник. Отв. редактор А.С.Герд. СПб.: Изд-во С.-Петербург. Ун-та, 1996. – С. 100-112.

3. В.Д.Буторов. Моделирование синтаксиса естественного языка // Прикладное языкознание: Учебник. Отв. редактор А.С.Герд. СПб.: Изд-во С.-Петербург. Ун-та, 1996. – С. 142-160.

4. В.В.Богданов. Моделирование семантики предложения // Прикладное языкознание: Учебник. Отв. редактор А.С.Герд. СПб.: Изд-во С.-Петербург. Ун-та, 1996. – С. 161-200.

ҒЫЛЫМИ-ТЕХНИКАЛЫҚ ЛЕКСИКОГРАФИЯ ЖӘНЕ АРНАУЛЫ СӨЗДІКТЕР ТҮРЛЕРІ

ҚР БҒМ ҒК А.Байтұрсынұлы атындағы Тіл білімі институты,
Алматы. Тілтаным (Языкознание). 2006, № 2, 146-153 бб.

Қазақ тілінде ғылыми-техникалық стиль және оған тән терминдік жүйе жаңадан ғана қалыптасып келеді деуге болады. Себебі техникалық ғылымдардың қазақ тіліндегі көрінісі немесе дерек көздері түрлі монографиялық зерттеулер, оқулықтар, мақалалар, баяндамалар, тезистер, патенттер т.б. сан-салалы еңбектер арқылы танылады. Ал бұл аталғандардың саны мен сапасы қазақ ғалымдарының әлі де болса арнайы көңіл бөлуін қажет етеді. Ана тілімізде жарық көре бастаған жаңа бастаманың нәтижесінде тілімізге жаңа ұғымдар мен техникалық жаңа атаулар келе бастағаны белгілі. Яғни қазақ тілінің жалпы лексикографиясы саласының жаңа бағыты ретінде ғылыми-техникалық лексикография саласы да өз отауын құра бастады деуге болады.

Үндіеуропа тілдерінде, соның ішінде орыс тілі бойынша аталған бағыттың өзіндік даму тарихы мен теориялық және практикалық тәжірибесі баршылық. Міне, осындай жайттарды ескере отырып, мақалада ғылыми-техникалық терминологиялық сөздіктерге немесе арнаулы сөздер сөздігін құрастыруға қойылатын кейбір талаптарға қысқаша тоқталмақпыз.

Арнаулы сөздіктерді құрастырудың теориясы мен практикасы ғылыми-техникалық лексикографияның зерттеу нысаны болып саналып жүр. Мұндай арнаулы сөздіктер арқылы баспа материалдары мен әр түрлі мекеме орындарының іс қағаздарында көрініс тапқан терминдер мен әр ғылымдар салаларында орын алатын номенклатуралық деректер жайлы түсініктеме мен анықтамалар алуға болады.

Арнаулы сөздіктер белгілі бір сала маманы мен электронды машиналар қажеттіктері үшін құрастырылады. Ғылыми сала маманына қатыстылар: біртілді және аударма терминологиялық сөздіктер, идентификаторлар сөздігі, терминдердің жиілік сөздіктерді, рубрикаторлар мен классификаторлар сөздіктері. Арнаулы сөздіктердің екінші тобына, яғни электронды машиналардың қажеттіктеріне – түрлі типті ақпаратты жүйелерге арналған ақпараттық-іздестіру тезаурустары [5] және мәтінді автоматты өңдеу жүйелерінде қолданыс табатын машиналық сөздіктердің [6] әртүрлі типтері (негіз сөздердің, тіркестердің, қосымшалардың) жатады.

Кез келген тілдің дәстүрлі түсіндірме сөздіктеріне қарағанда ғылыми-техникалық терминологиялық сөздіктер, ең алдымен, белгілі білім аясы мен техникалық саланың мамандарына арналатыны белгілі. Терминологиялық сөздіктер, өздерінің бағышталуына байланысты – жоғары білімді мамандар мен ғылыми әдебиеттерді аударумен айналысатын кәсіптік аудармашыларға арналып құрастырылады. Жоғары дәрежелі мамандарға мұндай сөздіктер тек қажетті деген сөзді тауып алуға ғана мүмкіндік туғызып қоймай, сонымен бірге оның ұғымын анықтауға және әр түрлі басылымдарда кездесетін кейбір синонимдік сөздер мағыналарын ажыратуға да мүмкіндік туғызады.

Ғылыми мақалаларды бір тілден екінші тілге аудару кезінде идеографикалық принципте құрылған біртілді және көптілді түсіндірмелі-аударма терминологиялық сөздіктер де аса қажет [2; 6].

Ғылыми-техникалық терминологиялық сөздіктерді орта және жоғары теникалық оқу орындарының студенттеріне, әр салалық өндіріс орындары мен ауылшаруашылық аясындағы сан-салалы мамандарына және радио мен баспа қызметкерлеріне және т.б. арналған сөздіктер деп те ажыратуға болады. Сол сияқты, мектеп оқушыларына арналған терминологиялық сөздіктер де жеке түрде шығарылып жүр. Бұл сөздіктердегі деректерді оқырманға ұсыну тәсілінде оқушының жас ерекшелігіне айырықша көңіл бөлу негізгі ұстаным болып табылады. Мысалы, терминологиялық сөздікте қызықтыру элементтерімен бірге суреттердің, сызбалардың орын алуы. Соңғы кездерде ғылыми-көпшіліктік терминологиялық сөздіктермен қатар бөлек ғылыми аяның терминдік жүйесін анықтайтын арнаулы ғылыми сөздіктер де жарық көре бастады. Әдетте, мұндай сөздіктердің құрылымы әліпби тәртібінен гөрі, жүйелілік принципке негізделіп жасалады.

Қолданылу аясына қарай, тіркейтін, түгелдейтін – «узус» сөздіктерін, белгілі ережелерге бағындыруды көздейтін – регламенттеуші сөздіктерін, оқу жұмысын көздейтін – мәлімет беру, жүйелендіруші терминологиялық сөздіктер де жарық көруде [3; 4].

Тіркеу міндетін атқаратын немесе «узус» атты сөздіктер қатарына терминологиялық сөздіктердің басым көпшілігінің қатысы бар. Ол сөздіктер сол кезеңге және қарастырылып отырған білім саласына тән терминологиялық стандартты анықтайды.

Регламенттеуші сөздіктер нормативті және нормативті емес деп екіге ажыратылады. Нормативті терминологиялық сөздіктерге айқын түрде мысал бола алатын сөздіктер – стандарт терминдер сөздігі мен ұсынылған (бекітілген) терминдердің сөздігі.

Стандарт терминдердің анықтамалары бойынша олар – ресми, мемлекеттік-құқықты статусқа ие болатын жүйелі типтегі терминологиялық сөздіктер. Терминологиялық стандарттардың құрамы – кіріспе, негізгі және әліпбилік көрсеткіштерінің бөліктерінен тұрады.

Кіріспе бөлікте стандарттың таратылу аясы, стандарттау нысанын анықтау және қолданылуға болмайтын терминдер, синонимдер мен аударма баламалары жайлы мәліметтер беріледі.

Стандарттың негізгі бөлімінде осы стандартқа қатысты қабылданған ұғымдар жүйесіне сәйкес орналасқан терминдер мен олардың анықтамалары қамтылады.

Нормативті терминологиялық сөздік, ол бір ғана терминдік жүйеге қатысты болуымен бірге терминдердің қолдану шарттылығы арнайы құжат арқылы бекітілуін қажет етеді. Сондықтан нормативті терминологиялық сөздіктерді құрастырудың маңыздылығы аса зор деп саналады.

Арнаулы ғылым-техникалық терминологиялық сөздіктерді құрастырудың алға қойған мақсатына қарай, олар – **ақпараттаушы** және **жүйелендіруші** деп екіге ажыратылады.

Арнаулы ақпараттаушы сөздікке мысал ретінде ақпараттық-іздестірудің тезаурустық түрін атауға болады. Ақпараттық-іздестіру тезаурусы (АІТ) деп біртілді қадағаланып отыратын және лексикалық бірліктерге өзгеріс енгізіп тұру сипатындағы сөздікті атайды. Бұл сөздік бір ғана тілдің лексикасына негізделіп, ақпаратты өңдеу мен іздестіру қажеттігі үшін құрастырылады. Ақпараттық-іздестіру тезаурусы (АІТ), ол – ақпаратты автоматты өңдеу мен іздестіруді жеңілдететін сөздік түрі. АІТ терминдер арасындағы мағыналық, семантикалық қатынасты да толық ашуға жәрдем етеді. АІТ-ның негізгі бірлігі – терминдер мен дескрипторлар. АІТ-ның әліпбиленген лексика-семантикалық бөлігі – бір жерге жинақталған дескрипторлық сөздік мақалалары. Әрбір дескрипторлық сөздік мақаласы термин-дескриптордан бастама алады да, ал оның төменінде – осы дескриптордың мақала шеңберіндегі синонимдері беріледі және негізгі дескриптормен грамматикалық немесе ассоциативті қатынаста болатын басқа да терминдер орналасады [5].

Жүйелендіруші арнаулы сөздіктерге жататындар: ***классификаторлар, рубрикаторлар, жиілікті сөздіктер*** және ***терминдік жүйе*** сөздіктері. Классификаторларда реттелген түрде ғылым мен техниканың жүздеген нысан атаулары және олардың сипаттамалары беріледі. Мұнда ең алдымен техника-экономикалық көрсеткіштердің классификаторларын, өндіріс пен ауылшаруашылық өнімдер атауларын және т.б. атауға болады. Осындай терминологиялық сөздіктер ішінен әсіресе рубрикатор-сөздіктер маңызды рөл атқарады. Рубрикатор-сөздіктерді сала мамандары мен мекемелер үшін ақпараттық қызмет ету жүйесіндегі әдістемелік құжат деп те атауға болады.

Рубрикаторлар, көбінде, өндірістік саланың негізгі бағыттарының жүйеленген ретін және маңызды нысандар мен процестерді қамтиды.

Салалық жиілік сөздіктердің өзіне тән маңыздылығы – кейбір бөлектелген терминдердің жиіліктерінің қамтылуы. Сөздіктің ең бір аса маңызды сипаты – терминдердің тақырыптық сипатта қамтылуы. Егер жиілік сөздік бірнеше пәндік аяны қамтитын болса, онда ол көпсалалық сөздік деп аталады. Көбіне, ғылымның бөлініп шығуынан тарсалалық сөздіктер жасау қажеттігі туындайды. Мұндай сөздіктерді жасау тәжірибесінде түсіндірме терминологиялық сөздік типтерін құрастыру да аса маңызды. Түсіндірме терминологиялық сөздіктердің міндеті – нақты бір сала бойынша анықтама құралы болу және барлық терминдерді қамтумен бірге ұғымдар жүйесінің көрінісін беру. Мұндай мәліметтер мамандардың ғылыми және тәжірибелік жұмыстары үшін аса қажет деп саналады.

Терминологиялық түсіндірме сөздікте ғылым мен техниканың дамуынан пайда болатын барлық жаңа ұғымдардың терминдік көрінісін берумен қатар, ол ұғымдардың анықтамалары да қамтылуы қажет.

Терминологиялық сөздіктердегі мәліметтерді реттеуде дәстүрлі әліпбилік және ұялы әдістермен бірге статистикалық әдіс те қолданыс табады. Яғни мәліметтерді реттеуде терминнің мәтіндегі жиі не сирек кездесу сипаты ескеріледі [1].

Әр түрлі тілдердің деректерін қамту жағынан терминологиялық сөздіктер біртілді және екітілді, үштілді және көптілді аударма сөздіктер деп бөлінеді [2; 6]. Түсіндірме сөздіктерді жасаудағы негізгі ұстаным – терминнің жаңа мағынаға ие болуы.

Терминологиялық сөздіктердің әр түрлі типтерінің негізгі дереккөзі ретінде жетекші ғалымдар мен маман-тәжірибешілердің қаламынан жарық көрген белгілі бір білім саласына қатысты жинақтап-қорытындылаушы монографиялар мен мақалалар материалдары алынады. Сонымен бірге, қажетті материалдардың қатарына жоғары оқу орындарының ірі оқулықтары да жатады. Әдетте, тек осындай оқулықтарда ғана ғылыми ұғымдар мен терминдердің аса айқын және нақты анықтамалары беріледі. Энциклопедиялар мен басқа да сөздіктерде берілген салалық терминдердің барлығы бірдей ғылыми-техникалық терминологиялық сөздіктерде орын алуы қажет.

Ғылыми-техникалық терминологиялық сөздіктердің лексикалық қорына қосымша ретінде реферативті журналдардағы кітап мазмұнының рефераттары, жаңалықтары, патенттер мен іс қағаздарының деректері де қамтылып алынады. Аса «қысылған» және шоғырланған түрде берілген мұндай мәліметтер негізгі терминдер жайлы жеткілікті түрде түсініктеме бере алады.

Сол сияқты, Ресейдегі ғылыми-техникалық әдебиеттер мен құжаттарды аудару орталығының жүйелі түрде шығарып тұратын «Тетради новых терминов» атты еңбектің материалдары да терминологиялық сөздіктер үшін құнды дереккөз болып табылады.

Неше түрлі құрал-саймандарды жөндеуге қатысты және т.б. салалық, ведомстволық нұсқаулар мен нұсқаулықтарда кездесетін терминдер мен номенклатуралық белгілеулердің терминологиялық сөздік жасаудағы атқаратын рөлі де айтарлықтай мәнді.

Сонымен, сөздік түзу кезіндегі әр түрлі дереккөздерден таңдалып алынған терминдер негізінде **терминологиялық сөзтізбе** құрастырылады. Бұндай сөзтізбелердің реестрлік бірлігіне жататындар: **сөз, сөз тіркес, қысқартулар, символдар және олардың тіркестері**. Әдетте терминологиялық сөздіктің сөзтізбесі таза интуитивті тұрғыда түзіледі, яғни әрбір сала мамандарының жеке көзқарасындағы негізгі және маңызды деп аталатын терминдері таңдалып алынады. Әрине мұндай әдіс ғылыми әдебиеттерді талдауға негізделмеген, тек қана субъективті пікірлік ұстаным екені белгілі.

Ғылыми-техникалық терминологиялық сөздікті құрастырудан бұрын, маман-ликсикограф ең алдымен баспаға дайындалып жатқан еңбектің **хронологиялық шеңберін** анықтап алуы қажет. Ғылымның жаңа салалары, мысалы есептеу техникасы, радиофизика, биофизика үшін мұндай мәселе онша өзекті болмауы да мүмкін, себебі ондай сөздікке енетін кез келген лексиканың уақыт мерзімі беріден басталады. Ал жаратылыстану ғылымдарының - физика, механика, математика, химия, биология, география, геология және көптеген қоғамдық ғылымдар салаларында мәселе басқаша.

Бұл ғылымдардың орысша мәтіндері өзінің тарихи дамуында 150-200 жылдай мәзгілді қамтиды. Осы себептерден терминологиялық сөздікте көрініс табатын **терминдердің хронологиялық тәртібі** маңызды мәселелер қатарына жатады.

Аталған мәселені тарихи-филологиялық және тарихи-ғылыми тұрғыда кең түрде қарастыратын болсақ, ол ертеректегі мәтіндерді дұрыстап оқи білу, оның мағынасын дұрыс түсіну және бір тілден екінші тілге дұрыс аудару мәселелеріне қатысты деуге болады. Ежелгі терминдердің де қазіргі әдебиеттерде қолданыс табатынын ескерсек, аталған мәселе қай уақытта да өз құндылығын жоймайды.

Кейбір ғылыми кітаптар мен мақалаларда өткен кезеңдердегі ғалымдардың зерттеу нәтижелерімен пікірталас болып жатады. Мысалы, XIX-XX ғасырлардағы ғалымдардың зерттеулеріндегі ұғымдар мен концепцияларына көңіл бөліну жағы, көбінде, филология, математика, биология, химия салаларына қатысты болады. Егер мұның бәрі тек қана тіл арқылы іске асатынын ескерсек, терминологиялық сөздіктерге лексикалық қорды таңдап алу кезеңдік сипатта болатынына және тіл туралы ғылым тарихының өзі де ғылым тарихымен ұштасып жататынына көз жеткізуге болады.

Мәселен, орыс тіліндегі ғылыми-биологиялық мәтіндердің тарихы мынадай кезеңдерге бөлінеді:

1) XVIII ғ. бастап, XIX ғ. 50-жылдар аралығы – орыс тіліндегі биологиялық мәтіндердің бастапқы қалыптасу кезеңі;

2) XIX ғ. 50-60-шы жылдары мен XX ғ. бастапқы кезеңі (10-20-шы жылдар);

3) XX ғ. 10-20-шы жылдары мен бүгінгі күнге дейінгі кезең.

Яғни биология саласы бойынша терминологиялық сөздік құру барысында XX ғ. алғашқы жылдарына дейінгі биологияға қатысты терминдерді қазіргі биология тіліне жатқызуға болмайды [3; 4].

Әрине, терминологиялық сөздікті пайдалану кезінде сол сөздік ішінен термин жайлы, тіпті оның тарихына дейінгі мәліметтер табылып жатса, әр түрлі сөздіктермен әрекет етудің қажеттігі болмас еді. Әрі тарихи тұрғыда терең хронологиялық қабатты қамтып, әрі қазіргі кездің терминдері бойынша анықтағыш рөлін атқаруға қатысты төмендегідей екі түрлі маңызды ой-пікір бар.

Біріншіден, ол – ғылыми білім саласының айтарлықтай күрделі және әрдайым құбылып тұру сипаты, ұғымдар иерархиясының күрделі эволюциясы, ұғымдардың мазмұны мен көлемінің үнемі өзгеру сипаты, ұғымдар арасындағы терминдердің тұрақсыз үлестірілуі, терминологиядағы синонимдік қатарлардың тұрақсыздығы.

Екіншіден – басқа тілдердің лексикалық параллельдерін ескере отырып, оқырманға қазіргі терминология жайлы молынша мәлімет беру жағдаяты. Сондықтан терминологияның даму тарихын қамтитын өзінше бір бөлек арнаулы сөздік болғаны жөн.

Қалай болғанда да терминологиялық сөздіктің хронологиялық төменгі

шегі жайлы мәселе әлі де болса өз шешімін таппай тұр. Төменгі хронологиялық шек XX ғ. 20-30-шы жылдары ма немесе XIX ғ. мен XX ғ. аралығы ма? Оны тек нақты білім және ғылым салаларының негізгі кезеңдеріндегі тілдің даму сипатына қарай ғана шешуге болады [3; 4].

Терминологиялық сөздіктердің сөзтізбесін қалыптастырғанда субъективті пікірді азайтудың бірден-бір жолы – нақты ғылым мен білім саласына тән логика-ұғымдық жүйені алдын ала модельдеу. Арнайы мәтіндерді логика-ұғымдық тұрғыда талдау тәсілі терминологиялық сөздіктерді жасаудағы алдын ала қойылатын қажетті шарт деуге болады.

Терминологиялық сөздіктердің құрылымы жайлы мынадай екі негізгі мәселені атауға болады:

1) терминологиялық сөздіктегі мәліметтерді қалыптастыру принциптері жайлы мәселе;

2) терминологиялық сөздіктегі жеке фрагменттердің құрылымы жайлы мәселе.

Бірінші мәселеге байланысты мынаны айтуға болады. Сөздік материалын ұйымдастыруға қатысты ғылыми-техникалық лексикографияда классикалық лексикографиядағы сияқты – әліпбилік, әліпбилік-ұялы, идеографиялық жалпы принциптер қарастырылады.

Терминологиялық сөздіктердің басым көпшілігі әліпби принципімен құрастырылатыны белгілі. Бұл ұстаным кез келген сөздікті пайдалану мен ондағы бірліктердің орналасу тәртібі жағынан көптен бері өзін-өзі ақтап келеді.

Осы айтылғандарға қарамай, мамандардың көптеген практикалық мұқтаждығынан аңғаратынымыз, ол терминологиялық сөздіктердегі терминдердің орналасуы тек әліпбилік немесе әліпбилік-ұялы принциптерге ғана сүйенбей, сонымен бірге логика-ұғымдық негізде, яғни ұғымдардың қатал ғылыми классификациясы бойынша идеографиялық принципте жасалуының қажеттігі. Бұл жердегі идеографикалық принципті қайсы бір ұғымды таңба арқылы белгілеп жазу деп түсіну керек. Мұндай принцип нақты ғылымның, білім саласының даму қисынына сәйкес келуімен бірге, терминді мағынасы бойынша сөздіктен тез табуға да мүмкіндік жасайды. Мәселен, нақты ғылымдар мамандарына, көбінде, қайсы бір ұғымның қалай және қандай сөзбен аталатыны немесе ол сөздің синонимдерінің қандай ұғымдармен байланысы бар екендігін білу қажет болады. Маманның зерттеу тәжірибесінде және аударма практикасында белгілі бір терминді тұрпаты жағынан ұқсас болып келетін ертеректердегі зерттеулердің терминдерімен сәйкестендіру қажеттігі де туындауы мүмкін [4].

Идеографикалық принципмен құрастырылған терминологиялық сөздіктің негізгі құрамы – сол жағындағы терминдердің әліпби тәртібімен берілуі ғана емес, ол вербальды түрде берілген және шартты немесе формалданған тұрпаттағы ұғым болуы қажет. Аталған ұғым сөздіктің жалпы логика-ұғымдық құрылымында қатаң түрде өз орнына ие болады. Мысалы, Роженің тезаурусында [Roget’s thesaurus…, 1978] сөздік мақаласының басы ұғымның коды арқылы (үш орынды сан) және осы ұғымды білдіретін тірек

сөзбен басталады. Тірек сөзден кейін сызықша және сөз табының белгісі қойылады, сосын сол сөздің синонимдері беріледі.

Доризайфтың неміс тілінің идеографиялық сөздігінде [Dorn Sejf, 1958, S. 211] әрбір сөздік мақаласы белгілі бір тақырыптық топтағы ұғымның нөмірімен басталып, әрі қарай құрылымдық тұрғыда бір-біріне жақын жатқан осы ұғымды білдіретін сөздің синонимдік баламалары берілген.

Ғылыми-техникалық терминологиялық сөздіктің жеке фрагменттерінің құрылымы. Лексикографияның жалпы теориясына сәйкес, сөздік мақаласы, көп жағдайда, мынадай бөліктерден тұрады [4, 297-280 бб.]:

1) сөздік мақаласының аты;

2) терминдердің грамматикалық сипаттамасы;

3) терминнің қолданысына байланысты стильдік реңкі мен пайда болуына қатысты сипаттамасы;

4) терминнің семантикалық сипаттамасы.

Әліпбилік принципте түзілген терминологиялық сөздік мақаласының құрылымы түсіндірме және энциклопедиялық сөздіктердің құрылымына ұқсас болып келеді. Терминологиялық сөздіктерде сөздік мақаласының аты, әдетте, әліпби тәртібімен орналасады да, олар тек термин сөздер ғана емес, сөз тіркестер, сөздер мен символдардың комбинациясы, аббревиатуралық қысқартулар болуы мүмкін.

Әліпбилік принциппен құрылған ғылыми-техникалық терминологиялық сөздіктерде терминнің семантикалық сипаттамасын беру оның мағынасының логикалық анықтамасы арқылы немесе аударма терминологиялық сөздіктердегі баламасы арқылы іске асады. Терминнің әрі шағын, әрі дәл логикалық анықтамасының берілуі, оның семантикалық құрылымын сипаттаудың ең бір тиімді жолы деп саналады.

Терминологиялық сөздіктерде терминдердің мағынасын анықтаудың ең бір дұрыс жолы, ол әрбір ұғымның типіне қарай, бірізділікті сақтай отыра, алдын-ала белгілі параметрлерге сүйену. Терминнің толық түрдегі лексикографикалық анықтамасын құруға қатысты мынадай талаптарды ұстаным еткен жөн: логика-ұғымдық сызбаға алдын ала талдау жасау, ұғымның шығу тегін анықтау, ұғымға тән маңызды белгілерді анықтау, осы ұғымның басқа ұғымдар аясы ішінен ерекше белгілерін анықтау.

Арнайы әдебиетті аудару ісімен айналысатын аудармашыға қажетті сөздің баламасын табумен қатар оның синонимдерін де іздеп табу қажет болады. Яғни терминді мазмұндық тұрғыда кеңірек танып, сөздікте көрсетілген сөз, шындығында да түпнұсқадағы ұғымды білдіру-білдірмеуіне көз жеткізу керек болады [2; 6].

Ғылыми-техникалық терминологиялық сөздіктерді құрастыруда терминнің анықтамасын тұжырымды етіп беру үшін мынадай ережелерді басшылыққа алған жөн саналады [4, 298-300 бб.]:

1. Термин арқылы аталатын барлық ұғымдар мен нысандар анықтама ішінде қамтылуы қажет және, керісінше, анықтамада сөз болатын барлық

ұғымдар мен нысандар терминнің атындағы ұғымдар мен нысандарға сәйкес келуі қажет.

2. Терминнің анықтамасында пайдаланылатын белгілер қарастырып отырған тақырыптық ая бойынша ең маңызды және мәнді болуы қажет.

3. Терминнің анықтамасында басқа бір терминнің анықтамасына сілтеме жасауға болмайды. Себебі, егер ол терминнің анықтамасында тағы да басқа бір терминнің анықтамасына жүгінсе, оның шегі табылмай, қайта оралып алғашқы терминге келуіміз де мүмкін.

4. Анықтамада пайдаланылған терминдер осы сөздік ішінде анықталған болуы немесе бұрыннан да белгілі және бірізді түсінігі болуы шарт. Егер анықтамада пайдаланылған терминдер іргелес жатқан басқа пәндерге қатысы болса немесе ол жалпыпәндік термин болса, онда арнайы қосымшаларға сілтеме жасалғаны жөн. Басқаша жағдайда ол терминнің түсінігін қандай әдебиеттен оқып-білуге болатыны айтылуы керек.

5. Терминнің нақты мағынасын ашатын анықтама берілуі үшін оның ғылыми ұғымдар жүйесіндегі алатын орнын жете түсіну аса маңызды.

6. Терминдердің анықтамалары, көптеген жағдайда, сөздермен қатар, математикалық және химиялық формулалар мен неше түрлі безендірілулер арқылы көрініс табады.

7. Әр елдің тілдеріндегі бір ғана терминнің мазмұнына қатысты анықтамалар бір-біріне сәйкес болмауы мүмкін. Сондықтан, терминнің анықтамасын берерде басқа бір сөздікке сүйенетін болсақ, оның қай елдің сөздігінен алынғаны көрсетілуі керек.

Біртілді немесе аударма терминологиялық сөздіктерде терминнің семантикалық сипаттамасы синонимия, антонимия, текті-түрлік және ассоциативтік семантикалық қатынастар арқылы толықтырылуына әрі кеңейтілуіне болады.

Ғылыми-техникалық терминологиялық сөздікте нақты білім, ғылым, техника салаларының терминдік жүйесі көрініс табуы қажет. Ғылыми мәтіннің семантикалық мазмұнын дәл ашу, сол сала маманының қатысуынсыз мүмкін емес екендігі белгілі. Сондықтан ұғымның түсінігін сипаттау сол ғылыми сала маманының ісі де, ал оны барынша дәл және ықшам түрде тұжырымдап оқырманға жеткізу тілші-лексикографтың үлесі. Яғни ғылыми-техникалық терминологиялық сөздіктің сапасы сала маманы мен тілші-лексикограф маманының бірлесе әрекет жасауының нәтижесі деуге болады.

Мақалада сөз болған тұжырымдар қазақ тілінің ғылыми-техникалық лексикографиясының дамуына оң әсерін тигізер деген үміттеміз.

Әдебиет

1. Алексеев П.М. Статистическая лексикография. М., 1975.
2. Берков В.П. Вопросы двуязычной лексикографии. М., 1973.
3. Герд А.С. Основы научно-технической лексикографии. Л., 1986.

4. Герд А.С. Научно-техническоя лексикография // Прикладное языкознание: Учебник. Отв. редактор А.С.Герд. СПб.: Изд-во С.-Петербург. Ун-та, 1996. – С. 287-307.
5. Тезаурус научно-технических терминов. М., 1972.
6. Убин И.И. Автоматический переводной словарь. М., 1989.

АВТОМАТТАНҒАН ОҚУ ЖҮЙЕСІНІҢ НЕГІЗГІ ҰСТАНЫМДАРЫ

ҚР БҒМ ҒК А. Байтұрсынұлы атындағы Тіл білімі
Институтының «Тілтаным» журналы. Алматы.
2006. №4. 186-191-бб.

Оқытуды компьютерлендіру ісі қазіргі кездегі педагогиканың ең бір күшті бетбұрысы деуге болады. Сонымен бірге бұл білім беру жүйесінің алдында тұрған күрделі мәселелерінің біріне айналуда. Тек оқытудың автоматтанған түрі ғана дәлелділік (уәжділік, мативация) сипатының жоғарғы деңгейінің негізінде оқып-үйрету әсерінің нәтижелігін түбегейлі өзгерте алады. Себебі білім беру барысының ескі (дәстүрлі) әдіс-тәсілдеріне жұмсалатын адам еңбегі (инемен құдық қазғандай) есептеу техникасына, яғни компьютерге жүктеледі. Сондықтан адам баласының бір рет қана атқаратын еңбегі оқып-үйренушінің қалауынша компьютер көмегімен сан рет қайталануына мүмкіндік туады. Әрине, оқытуды автоматтандырудан педагогтың еңбегін мүлде қажетсіз деген ұғым туындамауы керек. Бірақ автоматтандыру әдісі оқытушыны «қара жұмыстан» – шығармашылық әрекеттен босатып, оны қарапайым механикалық амалдарға ауыстырады. Ал мұндай жұмыстарды адам (мұғалім) әрекетімен салыстырғанда ЭЕМ (компьютер) «шаршамай-шалдықпай» зор нәтижелікпен және өте жылдам орындай алады.

Оқытуды автоматтандыру барысы, ең алдымен, автоматтанған оқу жүйесін (АОЖ) жан-жақты зерттеп, құрастырып алып, тек содан кейін ғана оны қолдануды қажет етеді. Көбінде АОЖ ұғымына (түсінігіне) оқытуға қажетті техникалық құралдар жиынтығы мен оқыту программаларының дестесін (пакетін) жатқызады. Дестедегі әрбір бағдарлама тек белгілі бір сабақты жүргізуге, яғни оқытушы мен оқушы арасындағы әрекеттесудің бір ғана сеансын өткізуге икемделген. Мұндай тәсіл біржақты сипатқа ие, себебі «оқу жүйесі» ұғымы мынаған саяды: нақты автоматтанған оқыту (үйрету) программалары оқу жоспарына сәйкес келуімен бірге АОЖ-ның ішкі құрылымдық көрсетімімен тұтастық кешенін құрайтын мағынада түсінік алуы керек. Мұндай тұтастықта қарастыру келесі екі негіздемеге сүйенеді: дидактикалық материалдың ортақтығы және оқытудың стратегиясы мен тактикасын автоматты түрде таңдау. Дидактикалық материал ретінде компьютерлік программаларды пайдаланатын белгілі пәндік саланың немесе кең мағынадағы білім базасының деректері қарастырылады. Екінші негіздік жайт – АОЖ шеңберінде әрекет ету кезіндегі үйренушілердің (обучаемые)

алдыңғы қызметінде жүзеге асырылған мәліметтері, яғни оқу қызметі нәтижелерінің деректер базасы.

Дей тұра, автоматты түрдегі даралық оқыту программасының да белгілі дәрежедегі құндылық жағын теріске шығаруға болмайды. Бірақ АОЖ құрамына енетін дара оқыту программасының нәтижелігі әр уақытта жоғары сатыда тұратынын ескеру кажет. Соңдықтан оңдай компьютерлік программалар даралық оқыту мен топтық оқыту процестерінің бір жүйе шеңберінде өзара әрекеттесу мүмкіндіктерін ескере құрастырылғаны жөн болады.

Жалпы жағдайда, АОЖ-ның құрамы мынадай негізгі компоненттерден тұруы керек [Жарков, 60-6.]:

1) оқытудың техникалық құралдар кешені;

2) оқыту программаларының дестесі;

3) пәндік ая бойынша білім базасы;

4) оқыту әсерінің нәтижелерін тіркейтін деректер базасы.

АОЖ әртүрлі пәндерді оқыту барысында пайдаланылуы мүмкін. АОЖ-ның қызметі сол жүйедегі программалардың оқушыға тигізетін әсері арқылы анықталады. Бұл программалардың негізгі қызметтеріне жататындар:

- оқытудың өзі (собственно обучение);
- жаттығу;
- бақылау;
- бағалау.

АОЖ-ның қызметін бұлайша бөліп құрастыру оқыту программаларын жобалау мен құрастыру кезінде аса маңызды: программа сценарийінің жалпы көрінісін, оқып-үйренушілердің диалог формасын (егер диалог алдын ала қарастырылса), сонымен бірге белгілі дәрежеде оқып-үйренушінің оқыту программамен өзара әрекеттесу сеансынан бұрынғы және кейінгі әрекеттерін анықтайды.

Автоматтанған оқыту жүйесінің негізгі ұстанымдарының бірі ретінде оның толық сипаты саналатын болғандықтан, яғни АОЖ-да аяқталған оқыту үзіндісін жүзеге асыру мүмкіндіктерінде оқыту программаларының үш түрі де (ақпараттау, жаттықтыру, бақылау) толық сипатта орын алуы қажет. Тек осы аталған талап түгелімен орындалған жағдайда ғана автоматты түрде оқыту жүйесі толықтық сипатқа ие болады. Бірақ тек бір ғана оқыту программасы шеңберінде әртүрлі қызметтерді қамту алгоритмдерге жүктелетін мәселелердің негізгі айырмашылығына айтарлықгай күмән туғызады.

Жоғарыда аталған үш түрлі программалар үшін дидактикалық материалдар пайдалану тәсілімен айырым тапқанымен, бірдей дәрежеде қамтылуы қажет. Ал бұл жайт дидактикалық материалдардың (программада пайдаланылатын мәліметтердің құрылымы) өзара тәуелсіз болуының бірден-бір маңызды кепілі. Білім базасындағы немесе әмбебап тілде жазылған көрінісіндегі бірыңғай деректер өздерінің түрліше атқаратын қызметіне қарай, программалардың орындалу барысында әрқилы жүзеге асуы мүмкін. Мұндағы «әмбебап тілде жазылған көрініс» дегенді, деректер мен оларды өңдеу әдістерін формалды түрге келтіретін құралдардың жиынтығы деп түсіну қажет. Бірдей деректер қызметінің программадағы айырым табуы адам еңбегін

үнемдеу ақиқаттығынан басқа, ол материалдарға (ұқсас және қайшылықсыз) өзгерістерді тек бір рет қана енгізуге мүмкіндік туғызады.

Ақпараттаушы (информирующие) программалар арқылы оқып-үйренушілер (обучаемые) дидактикалық материалдармен таныса алады, яғни ондай материалдардан тиісті ақпарат ала алады. Оқып-үйренушіге қажетті ақпарат беру тәсілі деректер жиынын айрықша безендіруді, яғни оларды көрнекі түрде экран бетінде орналастыруды, аса маңызды ақпаратты айрықша бөлектеп көрсетуді қажет етеді. Дұрысында, бір қалыптағы «жансыз» мәтіндік ақпаратты қажетті графиктер арқылы, компьютердің әуендік және басқа да құралдары арқылы оларға «жан бітіріп», оқып-үйренушінің жадында сақтау мүмкіндігін арттыру. Оқып-үйренушінің қалауынша бұрын танысқан ақпарат үзіндісіне қайта оралып, жаттығуына да жағдай жасалуы қажет.

Жаттығуға қатысты программалар өз құрылымына бөлек тәсілді қолдануды қажет етеді: оқып-үйренушіні білім базасына немесе оқу материалының ең қиын деген және біркелкі қате жіберілетін үзінділері бойынша түзілген жауаптар жүйесіне айрықша назар аударуға мүмкіндік туғызу. Сонымен бірге оқып-үйренуші мен программа әрекеті арасындағы (және керісінше) диалог жүргізілгені де жөн болады. Мұндай программаның орындалу барысында, оқып-үйренуші тарапынан кеткен қателерді атап көрсету, тіпті программаның күрделігіне қарай, ол қателердің не себепті орын алғанын да хабардар етіп отыру жөн саналады.

Бақылаушы программалардың жұмысы кезінде әрбір тапсырманың жауаптары деректер базасының оқыту барысы нәтижелерінде сақталып отыруы қажет. Сонымен бірге, ақиқат және кездейсоқ қателерді аса ұқыптылықпен ажырата білу керек. Бұл жайт лингвистикалық оқып-үйрету программаларында аса маңызды деп саналады. Себебі жаңылыс басылған сөздердің салдары күрделі қателерге әкеп соғуы мүмкін.

Егер программада **бағалау** қызметі жүзеге асатын болса, қателерді әрі қарай да талдау қажеттігі туындайды, яғни кеткен қатенің «салмақ» деңгейін анықтау және орындалған тапсырма нәтижелерін статистикалық өңдеуден өткізу қажет болады.

АОЖ-да бірнеше **қосымша функцияларды** (қызметтерді) бөліп қарастыруға болады. Ондай **қосымша** қызметтерге төмендегілерді жатқызуға болады [Жарков, 62-6.]:

- оқу қызметі нәтижелерін тіркеу. Егер АОЖ-дың құрамында осындай деректердің базасы орын алса, онда мұндай қызметтің маңыздылығы арта түседі және оқып-үйренушілердің алған нәтижелерін тіркеу жұмысының әсерлігі де барынша өседі;
- оқып-үйренушіні ақпараттау қызметі. Егер білім базасының жүйесі арқылы оқып-үйренуші өз сұранысына қажетті деген ақпаратты алуына мүмкіндік туса, бұл қызмет толық көлемде жүзеге асады;
- диалогты ендіру қызметі. Диалогты пайдалану үшін жан-жақты зерттелген, қажетті деген айтылымдарды құрастыру мен тани білетін табиғи тілге барынша жақын келетін **жасанды тіл** құрастыру қажет;

- оқып-үйренуге қажетті стратегия мен тактиканы таңдау қызметі, яғни жүйені оқып-үйренуге икемдеу қызметі. Мұндай икемдеу қызметі үшін оқыту әрекеті бойынша құрылған деректер базасын пайдалану орынды деп саналады.

Аталған қосымша фунцияларды оқыту программаларының түріне қарай үлестірілу сипаты да өзгеше болып келеді. Мәселен, нәтижелерді тіркеу қызметін жүзеге асыру тек бақылаушы программаларында ғана мүмкін, ал ақпарттау фунциясының мұндай программаларға қатысы жоқ деуге болады.

Оқыту жүйесін құрастыруға байланысты бірнеше қолданбалы мәселелерді қарастырып, осындай қызметтерге қажет деп саналатын бірнеше ұстанымдар (принциптер) төмендегідей [Жарков, 62-63-66.]:

1. Көптеген АОЖ-дың жоғары нәтижелігінің жеткіліксіздігі оқып-үйрену ісіне кешенді түрдегі қатынастың (подход) жоқтығы. Оқытудың кешенділік ұғымын ЭЕМ-ның (компьютердің) оқып-үйрену барысына толық араласу жағдайы деп түсіну қажет. Егер автоматты оқытудың компьютерлік программалары үзілді-кесілді сипатта болып, оқу курсының тек бірнеше бөлімдерін ғана қамтитын болса, ол нәтижелі бола бермейді.

2. Тілді оқытып-үйрету (обучение языку) **көпаспектілі** болуы қажет. Яғни оқыту программасын құрастыру кезінде оқып-үйрету мазмұнын құрайтын ақпаратты тілдің әр деңгейіне бірдей қатысты болуын мақсат етіп қойған жөн болады. Мұндай жағдайда оқыту барысының бірсарындылығы (монотондығы) азаяды. Осының нәтижесінде үйренушіде тек статикалық аспектідегі тіл жүйесі туралы түсінік қана қалыптасып қоймайды, сонымен бірге әртүрлі тілдік бірліктердің қызметі туралы да түсінік пайда болады. Мысалы, оқушы жаңа атаулы (номинативті) бірліктерді меңгеру үстінде, оларды бірден әртүрлі синтаксистік құрылымдарда да қолдануына мүмкіндігі бар. Осындай мәліметтермен бірге оқып-үйренуші атаулы бірліктердің валенттілік пен тіркесімдік қасиеттерін де сәйкес (адекватно) игере алады.

3. Оқыту барысының табиғилығы, біріншіден, үйретілетін тілдің ең ірі және жиі кездесетін құбылыстарын көрсете алатын оқу материалдарын мұқият түрде таңдаудан тұрады; екіншіден, үйренуші жағынан басқаруды қажет ететін программаларда (мысалы, білім базасымен әрекет жасау) компьютермен жұмыс істеуде арнайы дайындықсыз-ақ диалог жүргізуді ұйымдастыру қажет болады; үшіншіден, компьютерлік программамен жүзеге асатын диалог тілі табиғи тілге барынша жақын дәрежеде жүргізілуі керек.

Үшінші жағдай қайсыбір нақтылауды қажет етеді. Көптеген диалог жүргізу мақсатындағы жүйелерде ағылшын тілі немесе соған ұқсас келетін тілдер қолданылады. Бірақ тілдерді оқыту программаларында, әсіресе, жаңадан үйренушілер үшін ана тілінің негіз болғаны жөн. Үйренуге (үйретуге) қажетті тіл ретінде, көбінде, ағылшын тілінің алынатынын ескерсек, оның интерфейсі ретінде, яғни компьютер мен адам арасындағы диалогты жүзеге асыратын программалық құрал тілі ретінде қолдану үшін ағылшын тілін аздап болса да игеру қажет балады.

4)Табиғилық және вариативтілік принциптері бір-біріне іргелес жатады. Тіл жүйесі, сөйлеу тілінде әртүрлі тілдік құрылымдарды жүзеге асыру арқылы бір ғана немесе бір-біріне ұқсас келетін коммуникативтік міндеттерді

көптеген тәсілдер арқылы шешуге мүмкіндік тудырады. Олардың қайсыбірін тандап алу, оқу барысында бейнеленетін экстралингвистикалық факторларға негізделеді.

5)Оқып-үйренуді автоматтандыру оқу барысында үйренушінің қажеттілік, қызығушылық деңгейін арттыруға мүмкіндік туындатады. Бірақ мұндай уәждік (мотивациялық) жоғары деңгейіне оқу программасының коммуникативтік негізде (яғни машина мен адам арасындағы диалогты жүзеге асыру арқылы) және оның жағдаяттық бағытталуы (бейімделгіштігі) (яғни тапсырманың жағдаяттығымен санасу) арқылы ғана жетуге болады.

АОЖ-ды жобалау мен жете зерттеуде мынадай негізгі міндеттердің шешімін табу қажет болады:

- материалды таңдау және оны топтау (әр сабаққа, тақырыптарға үлестіру);
- диалогтың сценарийін жете зерттеу;
- сыртқы пішін көрінісін безендіру немесе программаның дизайны;
- жүйедегі өзара әрекеттікті ұйымдастыру (оқып-үйретуге кешенді тәсілді жүзеге асыру);
- «деректер көрсетімі тілін» жете зерттеу;
- пәндік сала және онымен әрекет етуге қажетті интерфейс бойынша деректер базасын құрастыру.

Осындай әрбір ішкі міндеттердің әртүрлі оқу программаларының типтеріне қатынасы бірдей емес.

АОЖ-ның функционалды өзегін оқыту программалары құрайды. Себебі, үйренушінің оқыту программамен әрекеттесу барысы, шындығында да автоматтанған оқу жүйесі арқылы жүзеге асады. Міне, осы әрекеттесу ғана оқыту әсерлігінің аса жоғары уәждігі (мотивация) мен нәтижелігін анықтайды.

Оқыту программалары әртүрлі негіздер жиыны бойынша топтастырылуы мүмкін. Олардың ең маңыздылары мыналар [Жарков, 64-66-66.]:

1) оқып-үйренушінің ЭЕМ-мен (компьютермен) тікелей әрекеттесуіне оқыту программасының мүмкіндік тудыруына қарай. Кең мағынада алғанда, оқытуға қатысты программалар түр-түрге бөлінеді, бірақ олардың бәрі бірдей оқыту қызметін атқаруға бейімделмеген. Мәселен, ондай программалар қосымша сипатта мынадай қызметтерді атқара алады: оқу барысы жайлы ақпараттарды статистикалық немесе басқаша өңдеу қызметі, «әкімшіліктің» атқаратын көптеген қызметтерін есептеу құралдарына жүктеу (сабақтардың күнделік кестесін құру, оқып-үйренушілер мен оқытушылардың қызметтерінің нәтижелерін тіркеу және оларды техника жадында сақтау. Әсіресе, мұғалімдердің көп уақыты мен күшін қажет ететін оқу бағдарламасын жасау әрекеті, есеп беру жұмысын атқару және т.б.

2) оқыту процесінің аяқталған көрінісін жүзеге асыру үшін компьютерлік программа қандай қызметтерді атқаруына қарай: бақылау, қажетті әрекет түрін көрсету, оқып-үйренушінің әрекетін бағалау және т.б. Аталған міндеттердің әрбіреуін шешетін алгоритм тек оның өзіне ғана тән ерекшелікке ие, сондықтан олар бір ғана алгоритмдік тәртіп бойынша жүзеге аспайды; әртүрлі программалар үшін ортақтық сипат, ол оқыту барысына ешбір қатысы болмайтын кейбір таза формалды әрекеттер (процедуралар). Мысалы,

терезелік интерфейс, жинақгауыгштағы деректермен алмасу процедуралары, қыстырма (заставка) және басқа да программаның безендіру құралдары, оқып-үйренушімен диалог ұйымдастыру құралдары және т.б.;

3) ұсынылатын оқу міндетінің типіне қарай: дағдылы (обычное), тапсырмаға ену (погружение); тапсырманың жағдаятына ену;

4) оқу программасында диалогтың орындалуына және оның формасына қарай. Шындығында, автоматтанған оқу жүйесінің уәждік сипатын күшейтетін жеткілікті түрдегі маңызды құрал ретінде диалогты алуға болады, бірақ ол тағы да бір аса маңызды қызметке ие: бұл – «программа – оқып-үйренуші» «кері байланысын» (обратная связь) жүзеге асыратын бірден-бір құрал. Осының негізінде қандай да болсын формадағы күрделі оқу программалары диалогсыз жүзеге аспайды. Оқып-үйренуші жағынан да программаның орындалу барысын басқару және оның даралық ерекшелігін де есепке алу жұмыстары диалогсыз мүмкін емес. Диалог ескерілмеген оқыту программаларымен жұмыс жүргізу кезінде оқып-үйренуші әбден енжарлық күйде қалуға мәжбүр болады.

Диалог қайсыбір табиғи тілдің (ана тілі немесе ағылшын тілі) арнайы құрастырылған бөлігіне қатысты немесе жасанды бұйрықтар жүйесі негізінде құрылуы мүмкін. Анығы, оқыту программаларында табиғи тілді пайдалану қай жағынан алғанда да аса құнды, бірақ жасанды интеллект деңгейіне көтерілу қажет болатындықтан, ол көптеген қиындықтарға әкеп соғады; сонымен бірге оқып-үйренушілер бұйрықтардың арнаулы тілін алдын ала біліп алуы қажет болады және мұндай тіл тек бірен-саран оқыту программаларында ғана қолданылып қоймай, оның бүкіл оқыту жүйесінде орын алғаны аса мәнді.

5) оқыту программасының орындалу барысын оқып-үйренушінің өзі басқара алатын мүмкіндігіне қарай. Бұл жерде сызықтық (линейное) және тармақты алгоритмдер туралы сөз болып отыр. Бірінші жағдайда, яғни сызықтық алгоритмдерде программаның сценарийі бойынша оқып-үйренушілердің алдын ала берілген бақыланылатын және ұқсас әсерлер реттілігінен ауытқуға ешбір рұхсат жоқ. Екінші жағдайда, яғни тармақты алгоритмдер жағдайында аталған реттілік оқып-үйренушінің қалауынша құбылып тұруына мүмкіндігі бар. Бұл жағдай оқып-үйренушінің оқу барысына қатысының белсенділігін арттырады.

6) оқып-үйренушінің даралық ерекшелігінің ескерілуіне қарай. Даралық ерекшелік дегенде оқушының өзіне ғана тән сапалық сипатынан гөрі, «оқып-үйренушінің деңгейіне» қатысты қайсыбір сандық сипаттама деп ұғынған жөн. Өзара айырым табатын деңгейлер саны барынша үлкен болуы да мүмкін (кейбір жүзеге асырылған АОЖ-да оның саны тоғызға дейін жетеді). Даралық ерекшелікті ескеру оқыту тәсілінің дербес болуын таңдаудан туындайды, яғни ол толығынан алғанда, оқу барысының нәтижелігін барынша ұлғайтуға мүмкіндік жасайды. Мұндай әр оқып-үйренушінің дара мүмкіндігін есепке алу, программаның іске қосылмай тұрған кезінде, яғни алдын ала және программаның орындалу барысында да ескеріледі. Соңғы жағдайда бір деңгейден екінші деңгейге өтуге мүмкіндік туады. Мұндай ауысу әрекеті аса қарапайым немесе аса күрделі тапсырманы таңдайтын оқып-үйренушінің

қалауы бойынша да жүзеге асуы мүмкін және сонымен бірге программаның орындалу барысында алдыңғы әрекеттердің ойдағыдай жүруін автоматты талдау арқылы да жүзеге асыруға болады. Деңгейді бағалауда субъективтілікті (өзіндік сипатты) жою тұрғысынан автоматты талдау тәсілін қолдану ең ұғымды деп саналады;

7) оқып-үйренушінің білім базасымен әрекет ету мүмкіндігіне қарай. Әрине, мұндай мүмкіндік білім базасымен әрекет жасауды жүзеге асыратын программаларды ғана сипаттай алады, яғни ақпараттаушы, оқып-үйретуші және жаттықтырушы программалар бақылау жұмысына жарамсыз болып саналады.

8) АОЖ-да программалардың орындалу нәтижелерінің тіркелу жағдайына қарай. Бұл мүмкіндік өзінің басқа да құндылықтарымен бірге нақты оқып-үйренуші мен АОЖ арасында өтетін сеанстық жұмыстар әсерінің сабақтастығын қамтамасыз ете алады. Мысалы, әрбір оқып-үйренушіге қатысты қайсыбір программаның орындалу жағдаяты тіркеледі. Ал бұл оқыту барысында кездесіп тұратын кейбір тақырыптың сырт қалуынан және қайталануынан сақтаңдырады. Сонымен бірге нәтижелерді тіркеу ісі оқып-үйренушінің білім деңгейін оқу барысында ескеріп, реттеп отыруға мүмкіндік тудырады.

9) дидактикалық материалдардың программа мәтінінде немесе оған тәуелсіз деректер берілісінде орын алуына қарай. Мүндай жағдайда әртүрлі деректер жиынтығымен бір программа арқылы әрекет етуге мүмкіндік туады. Осының нәтижесінде оқыту жүйесін программалауға жұмсалатын адам еңбегі мен оқу-үйрену материалдарына тиісті өңдеулер жүргізу уақыты үнемделеді. Осы айтылғандарға байланысты автоматты оқу жүйесінде (АОЖ-да) көптеген міндеттерді шеше алатын және икемділігі мол деректер көрсетімі тілін (язык представления данных) зерттеп құрастырудың маңыздылығы арта түсуде.

Әдебиет

1. Автоматизированные обучающие системы на базе ЭВМ / Под. ред. А.Ф.Чернявского. Минск, 1980.

2. *Жарков И.В.* Автоматизированные обучающие системы // Прикладное языкознание: Учебник / отв. ред. А.С.Герд. СПб:. Изд-во С.-Петербург, ун-та, 1996. 528 с. (59-67-66.).

3. *Носенко Э.Л.* ЭВМ в обучении иностранным языкам в вузе. М., 1988. 103 с.

4. *Пиотровский Р.Г.* Компьютеризация преподавания языка. Л., 1988,

5. *Рыбаков Ф.И., Соколова И.Н., Лоскутникова Л.В.* Программированное обучение с использованием ЭВМ. М., 1968.

6. *Савельев А.Я., Новиков В.А., Лобанов Ю.И.* Подготовка информации для автоматизированных обучающих систем. М., 1986. 174 с.

7. *Фадеев С.В.* ЭВМ в преподавании русского языка как иностранного. М., 1990, 81 с.

ТАБИҒИ ТІЛДІҢ МӘТІН МОРФОЛОГИЯСЫН

АВТОМАТТЫ ТҮРДЕ ТАЛДАУ МӘСЕЛЕСІ

ҚР БҒМ А.Байтұрсынұлы атындағы Тіл білімі институты.
«Мемлекеттік тіл саясаты: терминология, аударматану, ресми құжат тілі» атты халықаралық ғылыми-теориялық конференция материалдары. Алматы-2007. 141-144-бб.

«Компьютерлік лингвистиканың» айналысатын мәселелерінің ең маңыздыларының бірі – табиғи тілдің мәтін морфологиясын автоматты түрде талдау. Егер, дәстүрлі түсінік бойынша, «сөз морфологиясы» деген ұғым сөздің тұлғалық құрылымын, яғни сөз түбірін және оған жалғанатын қосымшалардың түрін анықтайтын болса, компьютерлік тіл біліміндегі «морфология» терминінен басқаша ұғым туындайды. Дәлірек айтсақ, ол сөздің мәтін бойындағы сыртқы пішініне қарай (тұрпатына қарай), тілдік құрылымның әр қилы деңгейлері бойынша тілдік мәліметтер алу мүмкіндігін білдіреді. Алғашында «морфологиялық талдау» ұғымы айтылған мағынада машиналық аударма саласында пайда болды [1].

«Морфологиялық талдаудың» жаңаша түсінігі бойынша алынатын ақпарат біздің дәстүрлі ұғымымыздағы морфологияға қатысы жоқ көптеген амалдардан тұруы мүмкін. Осының салдарынан компьютерлік тіл біліміндегі «морфологиялық талдау» ұғымы – амалдар ұғымы (операционное понятие), яғни сөздің сыртқы пішіні бойынша тілдік ақпараттарды танып-білу үшін жүргізілетін іс-әрекет деп ұғыну қажет. Басқаша айтсақ, егерде дәстүрлі тіл білімі ұғымы бойынша «морфологиялық талдау» кезінде «нені талдаймыз?» деп сұрақ қоятын болсақ, компьютерлік тіл білімінде «қалай талдаймыз?» деген сұрақ қойылады, яғни ол сөз тұлғасы бойынша кажетті тілдік ақпараттарды қай жолмен алуға болатындығымен айналысады.

Машиналық (компьютерлік) аударманың алғашқы тәжірибелерінде мұндай амалдардың саны көптеп кездесетін, ал бүгінгі таңда орыс және еуропа тілдері бойынша бұл мәселе біршама шешілді деуге болады [2].

Қазақ тілі мен басқа да түркі тілдері бойынша мұндай ізденістер айтарлықтай қолға алынбай жүр.

Аталған түсініктегі «морфологиялық талдау» бірнеше бағытта карастырылуда. Олардың біріншісі – сөзтұлғаны оның негізі мен оған жалгануы мүмкін болатын қосымшаларға ажырататын классикалық талдау негізінде модельдеу.

Екінші бағыт сөзтұлғалардың соңғы әріптерінің тіркесімдік зандылықтарының жиілік сөздіктердегі статистикалық мәліметтеріне сүйенеді.

Үшінші бағыт – соңғы кездердегі ізденістердің нәтижесі. Бұл бағыт бойынша теңдеудердің ашық жүйелерінің пішіні ретінде, морфологияның әмбебап математикалық моделі жасалады. Модельдің есептегіш мүмкіндігі неігізінде сөзтұлғаларды нормалау және қажетті грамматикалық ақпарат алумен қатар, сөзтұлғаларды жинақтау (синтездеу) да іске асады [3, 44].

Сөзтұлғалардың әріптік құрамының өзгеруін анықтайтын сөздерді өзара топтауды (сыныптарға бөлуді) негіз етіп, автоматтандырылған морфологиялық талдаудың алгоритмі құрастырылады. Мұндай топтау «морфологиялық топтау» деп аталып жүр [4]. Сөзтұлғалардың жазылуындағы әріптік өзгеріске ұшырау флективті және агглютинативті тілдерде біркелкі емес.

Мәселен, орыс тіліндегі: *«сижу – сидишь»* сөзтұлғаларының жазылуындағы сөз негізіндегі әріптер өзгеріске ұшыраса, сол сөздердің қазақша баламасы – *«отырмын – отырсың»* сөздерінің жазылуында косымшадағы әріптер өзгерген.

АМПАР атты машиналық аударма жүйесінде сөздердің синтаксистік қызметі мен септік және т.б. жалғаулардың негізінде сөздердің морфологиялық кластары екі түрге бөлініп қарастырылды:

1) негіздерінде өзгеріс болатын сөздер тобы,

2) сөздердің флективті тобы.

Соңғы топ белгілі сөздерге ғана тән қасиеттегі белгілер жүйесі арқылы немесе осындай қасиеттерді бойына сақтаған дерек сөздер арқылы сипатталады.

Енді морфологиялық талдаудың түрлеріне қысқаша тоқталайық. Олар төмендегідей:

– негіз сөздердің сөздігі арқылы морфологиялық талдау жүргізу;

– сөзтұлғалар сөздігі көмегімен морфологиялық талдауды іске асыру;

– логикалық көбейту әдісін қолданып, морфологиялық талдау жасау;

– морфологиялық талдауды сөздіктердің көмегінсіз (сөздіксіз) арнайы кестелер арқылы іске асыру.

Еуропа тілдерін зерттеуде көп тараған морфологиялық талдау түрі – негіз сөздердің сөздігі мен кейбір көмекші кестелер жүйесін пайдалану. Аталған сөздіктерде ішкі флексиясыз, жәй және күрделі сөздердің негіздері мен олардың негіз тұлғалары (формалары) толық беріліп отырады. Сөздіктегі әрбір негіз сөзге морфологиялық кластардың екі түрін ажырататын шартты белгі – «код» қойылады. Ал омонимдік негіздерге шартты белгілердің (кодтардың) тіркесімдік түрлері беріледі.

Морфологиялық талдаудың екінші түрі, яғни сөзтұлғалар сөздігі арқылы жүргізілетін талдау да зерттеушілер тәжірибесінде көп қолданыс тапқан, кең тараған талдау түрі болып саналады. Олай дейтініміз, морфологиялық таддаудың компьютерлік алгоритмінде морфемаларға бөлшектеу мен оларды сөздік бойынан іздестіру әрекеттері негіз сөздер сөздігіне қарағанда жеңіл іске асады. Бірақ талдаудың бұл түрінің өзіне тән осал жерлері де жоқ емес. Мысалы, егер іздестіретін сөз сөзтұлға сөздігінде кездеспеген жағдайда, біз оның грамматикалық ақпараты жайлы да ештеңе біле алмаймыз. Сондықтан сөзтұлға сөздігі көмегімен талдау жүргізу жүйесінде аффикстер мен түбір сөздер тізімдері берілуі қажет. Осымен бірге бірнеше сөзтұлғаларды бір лексикалық бірлікке сәйкестендіруге қажетті, оларға тән қасиеттерге ие, белгілер (атрибуттары) толық берілгені талап етіледі. Бұл айтылғандардан

туындайтын қорытынды – морфологиялық талдаудың негіз сөз сөздіктері арқылы жүргізілгені ұтымды.

Морфологиялық талдаудың «логикалық көбейту» әдісі зерттеу аясынан айрықша орын алады. Бұл әдісте «сөздік функциясы» деген ұғымға айрықша көңіл бөлген жөн. Әрбір «сөздік функциясына» сөзтұлға функциясы және оған тән ақпарат сәйкестендіріледі. Осының негізінде әр сөзтұлғаның өзіне ғана тән ақпараттық деректер арқылы функция мәнін арнайы кестемен беруге мүмкіндік туады. Бірақ аталған функцияның мәнін басқаша, төменде көрсетілетіндей тиімді амалдар арқылы беруге де мүмкіндік бар:

1) сөзтұлға – әріптер тізбегі ретінде морфемдік сегменттерге бөліктенеді;

2) сөзтұлға – морфемдік сегменттер тізбегі ретінде, басқа морфемдік элементтердің реттелмеген жиынымен ауыстырылады;

3) сөзтұлғаға, морфемдік жиын ретінде, белгілі бір ақпарат сәйкестендіріледі;

4) бұл ақпарат сөзтұлға жайындағы қажетті деген ең соңғы ақпарат түрінде қайтадан өзгертіліп беріледі [3, 46].

Әрбір морфемаға құрылымында сондай морфемасы бар сөзтұлғалардың жиынтық ақпаратын сәйкес қоюға болады. Логикалық түсінік бойынша мұндай ақпараттар жиынтығы дизъюнкциямен, яғни сөзтұлғаның құрылымдық белгілерінің бір-біріне қарсы қойылуымен сәйкес келеді. Сөзтұлға жайлы ақпарат осы сөзтұлғаның табиғатына тән қиылысудан немесе логикалық конъюнкция сияқты морфемалық акпараттардан тұратынына көз жеткізуге болады.

Морфологиялық талдаудың логикалық көбейту әдісі, көбінде, флективті тілдерде қолданылады және негіз сөздер сөздігінің болуын талап етеді.

5) қарастырып отырған автоматтандырылған морфологиялық талдаудың соңғы түрі – сөздіксіз талдау немесе «тәуелсіз» талдау. Ол аффикстер кестелері көмегімен және арнайы түзілген грамматикалық мағынасы жоқ сөздердің тізімі негізінде іске асады. Морфологиялық талдаудың бұл түрі зерттеу тәжірибесінде өте сирек қолданыс тауып жүр.

Енді автоматтандырылған морфологиялық талдаудың қазіргі күйі қандай деген сұрақтың жауабына қысқаша тоқталайық.

Қазіргі кезде кез келген мәтіннен компьютер арқылы сапалы ақпарат алу жүйесіне қойылатын талап заман сұранысына қарай өсті. Енді компьютер көмегімен морфологиядық талдау жасау алгоритмін іске косу мәселесіне, негізгілерін алғанда, мынадай талап қойылуда:

– кез келген тақырыптағы мәтіннің 98-99%-ын қамтитын көптақырыпты (политематический) «аса күшті» сөздік морфологиялық талдау жүйесінің негізін құруы тиісті;

– автоматты түрдегі талдау алгоритмі кез келген сөзтүрленістерін ескеретіндей мүмкіндікке ие болуы қажет. Осы жағдай ескерілсе ғана сөзтұлғаны жан-жақты тану мүмкіндігі артады және сан жағынан алғанда ондай мүмкіндіктің нәтижесі сөздікте қамтылған лексикалық бірліктен бірнеше есе артық болуы мүмкін;

– автоматтандырылған морфологиялық талдау жүйесінде «жаңа» сөздер сөздікте қамтылған басқа сөздермен бірдей дәрежеде қарастырылуы қажет және оларды дұрыс танудың ықтималдығы 90-95%-дан кем болмауы қажет;

– мәтінді компьютер арқылы өңдеу жылдамдығы тұтынушыны қанағаттандырарлықтай болуы қажет және оған енгізілетін мәтін көлеміне шек қойылмауы керек;

– компьютер арқылы жүргізілетін морфологиялық талдау жүйесінде қажеттікке қарай өзгерістер жүргізу (сөздікті толықтыру, өңдеу және т.б.) мүмкіндігі алдын ала ескерілуі қажет, яғни бұл аталған жүйені «оқытуға» барынша мүмкіндік тууы керек [3, 48-49].

Бүгінгі таңда, орыс тіліне және басқа да еуропа тілдеріне қатысты морфологиялық талдау жүйелеріндегі кейбір сөздіктер мен мәтіндердің көлемдері жайлы мынандай мәліметтерді айтуға болар еді: негіз сөздерден тұратын политематикалық сөздіктің көлемі 100 мың лексикалық бірліктен кем емес; сөздік көлемі 30 млн. сөзқолданыстан асатын политематикалық мәтін негізінде жасалғандықтан, оның кез келген тақырыптағы ғылыми-техникалық мәтінді қамту дәрежесі өте жоғары; көлемі 3 млн. сөзқолданыстан тұратын мәтіннің негізінде түзілген сөзтұлғалар сөздігі 46 мың лексикалық бірлікті құрайды. Сөз болып отырған екі сөздікте де сөздердің құрылымы мен синтаксистік қасиеті жөнінде толық түрдегі грамматикалық ақпарат берілген [5].

Автоматтандырылған морфологиялық талдау сөздіктердің айрықша (жаңа, арнайы) түрлерін өмірге келтірді. Оларда тілдік бірліктердің тұлғасына, туындау ережелеріне, сөзтудырғыш, сөзтүрлендіргіш топтарына қарай реттеу жақтары қарастырылды. Осындай типтегі сөздікке тілдік бірліктің соңғы жағынан бастап әліпбиге түсірілген кері әліпбилік (обратно-алфавитный) сөздік деп аталынатын сөздікті жатқызуға болады. Морфологиялық талдаудың талабына сай түзілген бұл сөздікте сөзтудырғыштық, сөзтүрлендіргіштік типтегі сөздер және соңғы қосымшалары бірдей болып келетін күрделі сөздер өзара жіктелген ретте көрініс табады. Кері әліпбилі сөздік бойынша сөздердің морфологиялық құрылымына қатысты көптеген мәселелерді шешуге болады. Мәселен, қосымшалары бірдей болып келетін сөздердің топталып берілуі олардың грамматикалық сипатын және қосымша мен қандай сөзтүрлендіргіш типке жататындығы арасындағы қатынастарға қатысты деректерді анықтауға мүмкіндік туады. Грамматикалық форманттардың синонимдік, омонимдік қатарларын жеңіл ажыратуға және олардың тіркесімдік, сандық мөлшерінің сипаттарын анықтауға мүмкіндіктер туындайды. Кері әліпбилі сөздік бойынша, аналитикалық тілдерге қарағанда флективті және агглютинативті тілдер жүйелерінен морфологиялық форманттарға байланысты ақпараттарды көптеп алуға болады.

Осы уақытқа дейінгі қазақстандық «Статистика тобының» қазақ тілін ЭЕМ-ның және компьютердің көмегімен лингво-статистикалық зерттеу барысындағы қол жеткізген көптеген ғылыми нәтижелерін, яғни әр мақсатпен түзілген жиілік сөздіктер мен олар арқылы алынған құнды

деректер мен түжырымдарды әрі қарайғы іс-тәжірибемізде дұрыс пайдалана білуіміз керек. Ең алдымен, сол ғылыми жетістігімізді үнді-еуропа тілдеріндегі «компьютерлік лингвистика» саласы бойынша орындалып жатқан ғылыми-зерттеу жұмыстарының деңгейіне дейін көтеру жолында жұмыс жасалу қажет.

Осы мақалада сөз болған компьютердің араласуымен ғана жүзеге асатын морфологиялық талдаулар мен басқа да зерттеу түрлерін мемлекеттік тілімізге пайдалана отырып, ана тіліміздің компьютерлік лингвистика саласын дамытуды және соның нәтижесінде қол жеткізген жетістігімізді күнделікті өмір қажетіне пайдалануды заман талабы деп түсінеміз.

Әдебиеттер:

1. *Василевский А.Л., Марчук Ю.Н.* Вычислительная лингвистика. Учебное пособие для студентов отделения прикладной лингвистики. М.: МГПИИЯ им. М.Тореза, 1970.
2. Лингвистические вопросы алгоритмической обработки сообщения. М.: Наука, 1983.
3. *Марчук Ю.Н.* Основы компьютерной липгвистики. Учебное пособие. М., 2000.
4. *Белоногов Г.Г., Новоселов А.П.* Автоматизация процессов накопления, поиска и обобщения информации. М.: Наука, 1979.
5. *Зеленков Ю.Г.* Морфологический анализ в системах автоматической обработки научно-технической информации. Канд. дисс. М.: ВИНИТИ, 1988. 145 с.

МАШИНАЛЫҚ АУДАРМАНЫҢ ҒЫЛЫМИ БАҒЫТ РЕТІНДЕГІ ДАМУ КЕЗЕҢДЕРІ

ҚР БҒМ А.Байтұрсынұлы атындағы Тіл білімі институтының «Қазақ тілінің лексикология, лексикография, фольклортану мен көркем аударма мәселелері: қалыптасуы, дамуы мен болашағы» атты халықаралық ғылыми-теориялық конференция материалдары. – Алматы. 2007. 241-246-бб.

Машиналық аударманың өмірге келуіне, біріншіден, XX ғасырдың екінші жартысынан бастап әр елдерде (континенттерде) бірнеше тілдегі ақпарат ағымының карқындап өсуі, екіншіден, ғылыми-техникалық прогресс үшін оларды меңгеру қажеттігіне қатысты әлеуметтік себептер негіз болды. Осыған байланысты машиналық (автоматты) аударма жұмысының даму барысына қатысты кейбір жайттарға тоқталайық [4].

Аталған әлеуметтік себептер аударма мәселесінің шешімін табуды және оның ұтымды жолдарын іздестіруді қажет етті. Бұған ең алдымен аудармашының тілдік қызметінің ерекшеліктері ішінен шеттілдік мәтіннің құрылымын зерттеудегі таза формалды талдау жолын жатқызуға болады [11]. Мысалы, аудармашыға мәтіннің грамматикасы және лексиканың көмекші сөздері мен есімдік сөздері және басқа да осы сияқты функционалды грамматиканың элементтері белгілі болғанымен, салалық мәтіннің арнаулы лексикасы таныс болмауы мүмкін.

Шеттілдік мәтінді аударудағы оның тұрпаты (формасы) бойынша мән-мағынаны ашу ерекшеліктерін модельдеу, ең алдымен, аудармашының іс-әрекетін зерттеу нысаны ретінде алуға бірден-бір себепші болды. Сонымен мұндай модельдеу шеттілдік мәтінді аударуда тілдің ақиқат деректерін «шеттетіп», тілдің семантикасын арнайы талдамай-ак, формалды тұрғыда аударма баламасын табуға мумкіндік жасайды.

Автоматты аударма жасаудың бастамасы ретінде ресми түрде 1949 жыл саналады. Тілді формалды тұрғыда қарастыра отырып, электронды-есептеу техникасының көмегімен аударманы жүзеге асырудан бұрын оны механикалық машиналар жолымен іске асыру идеясы аталған мезгілден де көп бұрын бастама алғаны белгілі. Мысалы, А.Кахердің ойлап тапқан аударма-жазба машинасы жайлы мәлімет «Уава таа» газетінде 1924 жылдың 24 ақпанында шағын хабарлама ретінде жарық көрген болатын [12]. 1933 жылы Кеңес Одағында П.П.Смирнов-Троянскийге бір тілден екінші тілге аудару машинасын ойлап шығарғаны үшін арнайы патент тапсырылғаны да жұртшылыққа мәлім [13].

Аударма жасауда екі немесе одан да көп тілдердің бірліктері арасындағы мағына сәйкестігін анықтау үшін алғашында механикалық тәсілдер колданылды. Сондықтан аударма жұмыстарының жетістіктері әлемдік ғылымға тек XX ғ. 50-жылдарында ғана белгілі бола бастады. Ең алғашқы машиналық аударма жүйелері АҚШ-та пайда болды. Тарихи тұрғыда қарастырғанда машиналық аударма жүйесін құрастыруға жалпытілдік көзқарастағы криптография (шартты белгілермен жазылған құпия жазу) саласының теориясы мен практикасы негіз болды.

Екінші дүниежүзілік соғыс кезінде әр елдерде, әсіресе Англия мен АКШ-та құпия хаттардағы шифрлы жазуды үйреншікті жазуға ауыстыру (дешифровка) жұмысын машинасыз да және машинамен де әрекет ету арқылы көптеген нәтижелерге қол жеткені белгілі. Мысалы, гитлерлік басқару орындарының шифрланған бұйрықтары мен хабарларын жай жазуға айналдыратын «Ультра» атты дешифровалды машинаның құрастырылғанын айта кетуге болады.

Лингвистикалық тұрғыда қарастырғанда, шифрланған (кодталған) хабарды дәстүрлі жазуға келтіру дегеніміз, ол лексикалық қоры белгісіз және грамматикалық ережелері әлі анықталмаған шет тілінде жазылған мәтінді аудару жағдаятымен барабар деуге болады.

Мұндай аударма жұмысы бір кодтағы жүйеден екінші кодтағы жүйеге ауыстыру әрекетімен бірдей келеді.

Маман-дешифровщиктердің пайымдауынша, шифрланған (кодталған) хабарды жазуға келтіру әрекетіне қарағанда бір табиғи тілдегі мәтінді екінші табиғи тілге аудару әрекеті жеңіл деп бағаланады. Себебі, қазіргі табиғи тілдерді сипаттайтын көптеген сөздіктер мен грамматикалардың негізінде тілдің өзіне ғана тән лексикалық, морфологиялық және синтаксистік ерекшеліктері алдын ала белгілі деуге болады. Ал есептеу техникасы ондай сөздіктер мен грамматикаларды қалаған тәртіпте, бағдарламаға сай өз жадында орналастыра алады.

1946-1947 жылдары ЭЕМ аркылы машиналық аударма жасау идеясы Принстонның ғылыми-зерттеу институтында арнайы мәселе ретінде талқыланады (У.Уивер, А.Л.Бут және т.б.). Ең алғаш бұл идея дешифровка ісінің американдық маманы Уоррен Уивердін меморандумында бірізді шешімге келтірілді. 1949 жылдың 15 шілдесінде атақты ғалым өз меморандумында көтерілген мәселелерді тіл және есептеу техникасы мен дешифровка салаларындағы екі жүзден астам мамандарға жария етті.

У.Уивер машиналық аударманың принциптік мүмкіндігінен мүлде бөлек тілдердің ұқсас қасиеттерінен және әр тілдің өзіне ғана тән формалды сипатта болуымен байланыстырады. Қорыта айтқанда, У.Уивердің пікірі бойынша, қытайша жазылған кітапты қытайша шифрланған (кодталған) ағылшынша немесе басқа ұлттық тілдердегі кітап деп түсінген жөн [9, 33-45].

Сонымен машиналық аударманың өзінше бөлек ғылыми бағыт ретіндегі даму кезеңі Уивердің меморандумынан бастама алады. Бұл бағытты, шартты түрде, негізгі үш кезенге бөліп қарастыруға болады [14; 5; 8, 91; *2,* 4-5]. Мысалы, ол хронологиялық жағынан былайша айырым табады: 1-кезең: 1949 жылдан бастап 60-жылдардың ортасына дейін, 2-кезең: 60-жылдардың ортасынан 70-жылдардың аяғына дейін; 3-кезең: 80-жылдардан қазіргі кезеңге дейін.

Мұндай үш кезеңге бөлуге негіз болған ғылым мен техниканың дамуындағы қол жеткен нәтижелер екені белгілі. Міне, солардың ішінен лингвистикалық, математикалық, техникалық машиналық аударма жүйесін жабдықтау мәселелері айрықша орын алады.

У.Уивердің меморандумымен жарияланған бірінші кезең 1949 жылы Вашингтон университетінде құрылған ғылыми орталықтың зерттеулерінен (ағылшынша-орысша машиналық аударма жүйесі) бастама алады. 1950 жылы А.Д.Буттың ғылыми жетекшілік етуімен французша-ағылшынша аударма орталығы құрылады. Кеңес Одағында 1955 жылы Мәскеудегі КСРО ҒА-ның В.А.Стеклов атындағы Математикалық институтында О.С.Кулагинаның жетекшілігімен ең бірінші машиналық аударма орталығы ашылады.

Осы кезеңде ең алғашқы зерттеушілерді үміттендіретіндей нәтижелер алынып, алғашқы машиналық тәжірибелер жүргізілген болатын. Мысалы. 1954 жылы АҚШ-та Джорджтаун университетінде проф. Леон Е.Достердің жетекшілігімен GAT жүйесі негізінде құрастырылған орыс тілінен ағылшын тіліне аударатын машиналық аударма жүйесі жұртшылық назарына ұсынылды. Бұл бір тілден екінші тілге ЭЕМ (IBM-701) арқылы орындалған

ең алғашқы аударма тәжірибесі болып саналады. Ғылымда бұл тәжірибе Джорджтауындық эксперимент деген атпен тарихта қалды [9, 171]. Аталған экспериментте «тікелей» машиналық аударма жағдаяты (бір тілдік жұпқа қатысты жүйелер) тексерістен өткен болатын. Жүйе сөздігінің көлемі бар болғаны 250 сөзді ғана қамтыса, ал оның грамматикасы тек алты синтаксистік ережеге ғана негізделеді [15, 22-23]. Кеңес Одағында француз тілінен орысшаға аударатын машиналық аударманың бірінші эксперименттік жүйесі О.С.Кулагина мен И.А.Мельчуктың авторлығымен 1955-1956 жылдары құрылғаны осы сала мамандарына белгілі [6]. Бұл жүйедегі негіз сөздер сөздігі 1236 сөз, ал сөз айналым сөздігі 250 бірлікті қамтиды. Аударма бірлігі ретінде толық сөйлем түрі есептеледі. Морфологиялық талдау алгоритмі жалғаулардың кестесімен әрекет ету арқылы аударылатын фразаның (сөйлемнің) сөздеріне қажетті деген грамматикалық акпаратты тіркейді. Бұл жүйенің артықшылығы – омонимдік сөздерді ажырата алатын топтамасы (блогі) болуында. Орыс тіліндегі фразаны синтездеу топтамасында сөздердің орналасу тәртібінің дұрыстығын тексеру мүмкіндігі алдын ала қарастырылған. Сонымен бірге, ТН.Молошнаяның жетекшілігімен ағылшынша-орысша аударма жасау жүйесінің алғашқы жобасы ұсынылды [10].

Жаңа бағыттың қарқынды дамуы көптеген мәртебелі ғылыми конференцияларда, симпозиумдарда, жарияланымдарда көрініс тапты және машиналық аударма проблемасымен айналысатын бірнеше ірі орталықтар, топтар мен зертханалар ашылуына себепші болды. Өткен ғасырдың 60-жылдарының басында олардың әр елдегі сандық көрсеткіші мынадай: ФРГ – 22, АҚШ – 21, КССРО – 10, КНР мен Жапония – 6-дан, Франция – 5, Чехославакия – 4, Великобритания мен Венгрия – 3-тен, ГДР мен Италия – 2-ден. Осындай топтар Бельгияда, Болгарияда, Израильде, Канадада, Люксембургте және т.б. елдерде де пайда бола бастады [1, 509].

Машиналық аударма жұмысының бастамасы жалпы алғанда өрлеу сипатында болды – тілшілер зерттеудің жаңа бағытын ашты, ал программистер күн тәртібіндегі күрделі және аса қызықты мәселелердің шешімін табуға қызыға кірісті. Бұл кезеңде тіл мен аударма процесіне қатысты барлық мәліметтерге қол жеткендей сезінілді. Сондықтан егер оларды формалды пішінге келтіру ісі жүзеге асырылса, машиналық аударманың әмбебаптық жүйесі құрылып, алға қойылған негізгі мақсатқа тез арада жететіндей көрінген болатын. Бірақ бастапқы және кейінгі тәжірибелердің нәтижелері бұл мәселенің шешімі ойлағандай оңай табылмайтынын байқатты.

Солай бола тұра, атқарылған жұмыстар тек машиналық аударма аясындағы зерттеулерге ғана емес, сонымен бірге ғылыми ізденушілерді тілдің теориялық және құрылымдық жағына да көңіл аударуға ынталандырды.

Қорыта айтқанда, автоматты аударманың бірінші кезеңі белгілі дәрежеде айтарлықтай нәтижелі болды. Себебі аталған кезеңде көптеген іргелі мәселелердің бастамасы анықталып, зерттеудің болашағы сараланып,

жолдары айқындалды. Мысалы, автоматты сөздіктерді құрастыру, аралық тілді құрастыру, сөздердің омонимдік сипаты мен синтаксистік құрылымы, яғни сөздердің сөйлемдегі тіркесімі, фразеологизмдердің жасалу жолдары және көптеген басқа да мәселелердің сыр-сипаттары анықталды.

Осы мезгілдерде, бүгінгі таңда да аса маңызды деп саналатын, синтаксистік талдаудың алгоритмінің алғашқы нұсқалары зерттелді: синтаксистік тәуелділік пен тікелей құрастырушыларды, болжамдық талдауды, көпвариантты синтаксистік талдауды, «фулькрумдер» әдісімен талдауды, конфигурациялық талдауды және т.б. әдіс-тәсілдерді ескере отырып, талдау жүргізу жақтары қарастырылды.

Машиналық аударма жүйелерін құрастыру аясындағы жүргізіліп жатқан жұмыстардың белсенділігіне қарамай, бірінші кезеңде зерттеушілердің ойлағанындай практикалық нәтиже алынбады. Аударманың тұтынушы сұранысын қанағаттандыра алатын өндірістік жүйесін құру жұмысының нәтижесіз болуының үлкен зардабы байқалды. Мысалы, зерттеушілердің бір бөлігі машиналық аударма мәселесін шешуден бас тартты. Осы бағытта жұмыс істейтін бірнеше топтың зерттеу әрекеттерінің тоқырап қалуына себепші болған АҚШ-тың Ұлттық Ғылым академиясының арнайы комиссиясының әуе соғыс күштерінің машиналық аударма жайлы жасаған талдау қорытындысын бұрмалап түсінудің (экстраполяция) де өз рөлі болды. Осының зардабынан көптеген ғылыми ұжымдар ұйымдастыру мен қаржылық қиындықтарға әкелді [5, 14].

Мұндай жағдайға душар болу кейбір ғалымдардың скепсистік көзқарасы да, машиналық аударма идеясының жүзеге асу мүмкіндігіне сенімсіздік білдіруі де, әсіресе, ғылыми-техникалық мәтіндерді аудару мүмкіндігіне күмән туғызуы да себепші болды. Ал көркем әдебиет мәтіндері, оның ішінде, әсіресе поэзия мәтініне байланысты машиналық аударма жасауға ең басынан-ақ ғалымдар қарсылық пікір білдірген болатын. Оған шығарманың көркемдік, бейнелік жағын ұстаным етудегі кейбір тілдік заңдылыққа қайшы келу жақтары негізгі себеп болды.

Осындай жағдайға байланысты И.Бар-Хиллелдің көзқарасына тоқтала кетуге болады. Ғалым жоғары сапалы толық автоматтандырылған машиналық аударма мүмкін емес және оны утопия деп санаған. И.Бар-Хиллелдің пікірінше, машиналық аударма кезіндегі негізгі кедергі сөздердің көпмағыналығы және тіл табиғатында ондай екімәнділікті ажырата алатын формалды көрсеткіштердің болмауы. Бұл қиындықтардың шешімі тек әлем туралы білімде жатыр деп тұжырым жасайды [3].

Жоғарыда айтылған сенімсіздік пікірлер әлем бойынша машиналық аударма мәселесімен айналысатын топтар санын азайтуға әкеліп соқтырды, екінші кезеңнің аяғында 100 топтан 38 топқа дейін азайғанын айтуға болады [5, 16-18].

Бірақ машиналық аударманың болашағына сенімді, дәйекті көзқарастағы ғалымдар қажымай, шаршамай-шалдықпай жоспарлаған ғылыми ізденістерін жалғастыра бергені сала мамандарына мәлім. Енді машиналық аударманың екінші кезеңі басталды. Машиналық аударма

мамандары зерттеудің теориялық жағын тереңдете түсіп. формалданған лингвистикалық модельдер жасауға, жүйелі және функционалды грамматиканы жазу мәселесіне кірісті. Бұл кезеңде семантиканы модельдеу жұмысына арнайы назар аударылды. Бірінші кезеңнің нәтижелері сараланып. болашақтағы прагматикаға бағышталған жүйе құрудың мүмкіндігі анықталды.

Екінші кезенде машиналық аударма мәселесінің зерттелуі өз жалғасын тауып, оның түсініктеме берілу жағына көбірек көңіл бөлінді, машиналық аударма мәселесі жеке тілдердің грамматикалық құрылымының азды-көпті толық түрдегі сипаттамаларында да қолданыс тапты. Сонымен бірге, олар автоматты талдаулардың іргелі алгоритмдері бойынша да сыннан өтті деуге болады.

Екінші кезеңге қатысты мынадай жайттарды атап кетейік: семантикалық талдау принциптерінің бұрынғыдан да қарқынды түрде зерттелуі және жоғары сапалы аударма жасауға қатысты талдаулар жасау жүйелерінде семантиканың қатысынсыз мәселенің шешілмейтіндігін толық мойындау. Машиналық аударма жүйесін құруға байланысты жұмыстардың екінші кезеңде жалғасуымен бірге, ол кезең аударманы автоматтандыруды тәжірибелік сыннан өткізу және оны өндіріске енгізу әрекеттерімен қоса сипатталады.

Машиналық аударамен айналысатын Ресейлік топтардың сан жағынан аз болуына қарамай, сөз болып отырған екінші кезеңде олардың зерттеу жұмыстарын аса нәтижелі деп сипаттауға болады:

а) лингвистикалық және математикалық жабдықтау тұрғысынан алғанда, бастапқы машиналық аударма жүйелерінің жақсартылған түрлерінің құрастырылуы;

ә) жаңаша типті жүйелердің құрастырылуы (Бүкілодақтық аударма орталығы, А.И.Герцен атындағы ЛГПИ-дегі «Статистика речи» тобы);

б) автоматтанған талдау жүргізетін жүйелердің қарқынды түрде зерттелуі;

в) машиналық аударма жүйелерін әртүрлі мәндік дәрежеге қатысты тәжірибелік сынау деңгейіне дейін жеткізілуі және бұрын-соңды құрастырылған жүйелердің жетістіктері мен кемшіліктері бірдей дәрежеде сыннан өту нәтижелері осыны байқатады.

Бүкілодақтық аударма орталығы жұмысының екінші кезеңінде бірнеше аударма жүйелері құрастырылды. Олардың ішінен тек екеуін ғана атап өтейік: Н.Н.Леонтьева жетекшілік еткен ФРАП атты французша-орысша автоматты аударма жүйесі [7] және Ю.Н.Марчук жетекшілік еткен АМПАР атты ағылшынша-орысша автоматты жүйе [8].

Сонымен машиналық аударманың даму кезеңінің екінші сатысы бірінші буынға қатысты практикалық жүйелердің дамуымен, тәжірибе жинақтаумен, оны іске қосумен, сөздік базасының кең түрде өріс алуымен және тәжірибелік тұрғыда болса да барынша жабдықталған теориялық жүйелерді құрумен сипатталады.

Қол жеткен практикалық және теориялық нәтижелер машиналық аударманың даму сатысының келесі үшінші кезеңіне көшуге даяр екендігінің айғағы болып саналды. Бірақ келесі сатыға өту жолы бірнеше экстралингвистикалық деректерді алдын ала анықтап алуды қажет еткені белгілі.

Біріншіден, 70-ші жылдардың аяғында ақпарат ағымы бұрын-соңды болмаған қаркында дамиды. Мысалы, американдық және канадалық корпорациялар бір жылдың ішінде 2.5 млрд. сөзге дейінгі мәтіндерді аударып отыруға мүмкіндіктері болған. Егер АҚШ-та 1966 жылы аударма жұмысының біржылдық шығыны 22 млн. доллар болса. 1981 жылы ол шығын 200 млн. долларға дейін өседі. Сол себепті қазіргі кезде автоматтандыру арқылы техникалық стильдегі мәтіндерді аударуды жеңілдету мен жетілдіру маңызды мәселеге айналып отыр.

Екіншіден, спеллер типті өндірістік лингвистикалық процессорларды құрастыру мен оларды баспа мен лексикографиялық жұмыстарда іс жүзінде қолдану, сол сияқты сөздерді өңдеу жүйесінде (word processors) күшті терминологиялық банктерді пайдалану – «лингвистикалық» автоматтанған жүйелерді қолдану арқылы нақты практикалық нәтижелерге ие болу мүмкіндігінің кепілі.

Үшіншіден, машиналық аударманың жаңа кезеңіне өтуге қатысты ерекше жасалған алғы шарт ретінде есептеу техникасының дамуына байланысты барынша жағдай жасалуы – дисплейлердің пайда болуы, дербес компьютерлер мен сканерлердің өмірге келуі, дербес ЭЕМ-нің күш-қуатының өсуі және т.б. Міне, осының бәрі теоретика-лингвистикалық зерттеулерге көзқарасты басқаша өзгертіп, оны практикалық деңгейге дейін көтерді.

Машиналық аударма жүйелерін құрудың бірінші және екінші кезеңдерін сипаттауда, біз нақты лингвистикалық ұжымдардың жұмыстарына айырықша назар аударды. Ал енді машиналық аударманың қазіргі кезеңін сипаттау үшін оның стратегиясын таңдау мен практикалық жүйелерін тиімді пайдалануға байланысты негізгі теоретикалық және техникалық жайттарды анықтауға қатысты мәселелер жеке қарастыруды қажет етеді. Әрине, машиналық аударманың мұндай негізгі мәселелерін талқылау машиналық аударма жүйесінің болашақтағы даму жағдаятын да сөз етуді керек етеді. Машиналық аударма жүйелерінің қазіргі сатыдағы даму ерекшеліктерін қарастырғанда оны әр кезеңге бөліп топтап, баға беру дұрыс бола бермейді. Себебі машиналық аударма қатысты ондай топтаулар есептеу машиналарының (компьютерлердің) кезеңді өзгерістеріне және мүмкіндіктерінің артуына да тікелей байланысты болады. Аталған жайттарды бір мақала көлемінде жан-жақты қамту мүкін де емес.

Қорыта айтқанда, ақпарат ағымының жаңа толқыны машиналық аударма жүйелерін күнделікті өмірде кең түрде қолдану қажеттігін айқындап отыр. Сондықтан осы саладағы болашақ ізденістер барысында өткен кезеңдерде қол жеткен жетістіктер мен кемшіліктердің барлығы бірдей ескерілуі қажет болады.

Әдебиеттер

1. Автоматический перевод. 1949-1963.: Критико-библиографический справочник / Сост. И.А. Мельчук, Р.Д-Равич. М., 1967. С. 516.

2. *Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л.* и др. Лингвистическое обеспечение системы «ЭТАП - 2». М., 1989. С. 295.

3. *Бар-Хиллел И.* Будущее машинного перевода // Науч. докл. высшей школы. Филол. науки. 1962. №4. С. 203-206.

4. *Беляева Л.Н., Откупщикова М.И..* Автоматический (машинный) перевод // Прикладное языкознание: Учебник. Отв. ред. А.С. Герд. СПб.: Издво С. Петербург. ун-та. 1996. С. 528.

5. *Кулагина О.С.* Исследования по машинному переводу. М.,1979. С. 319.

6. *Кулагина О.С., Мельчук И.А.* Машинный перевод с французского языка на русский // Вопр. Языкозн. №5, 1956. С. 111-121.

7. *Леонтьева Н. Н.* Система французско-русского перевода (ФРАП): лингвистические решения, состав, реализация // Машинный перевод и прикладная лингвистика: Проблемы создания системы автоматического перевода. Вып. 271. М., 1987. С. 6-26.

8. *Марчук Ю.Н.* Проблемы машинного перевода. М, 1983.

9. Машинный перевод: Сб. статей. М.. 1957. С. 314.

10. *Молошная Т. Н.* Алгоритм машинного перевода с английского языка на русскии // Проблемы кибернетики. Вып 3. М., 1960. С. 209-272.

11. *Панов Д. Ю.* Автоматический перевод. 2-е изд. М., 1958, С. 72.

12. Переводная машина в Таллине в 1924 г. // Сообщения по машинному переводу. Сб. 1. Таллин, 1962, С. 95-96.

13. Переводная машина П. П. Траянского: Сб. матер. о переводной машине для перевода с одного языка на другие, предложенной П.П.Троянским в I933 г. / Отв. ред. Д. Ю. Панов, М., 1959. С. 52.

14. Hutchins W.J. Transiation: Past, Present, Future. Harwood, 1986.

15. Machine Transiation Theoretical and Methodoiogikai Issues // Ed. By S. Nirenburg. New York, 1987, 350 p.

БАСПА ІСІН АВТОМАТТАНДЫРУ ӘДІСІНДЕГІ ЛИНГВИСТИКАЛЫҚ АСПЕКТІЛЕР

І.Кеңесбаев, М.Балақаев, Ғ.Мұсабаевтың 100 жылдығына орай 2007 жылы қараша айының 17-18 күндері А.Байтұрсынұлы атындағы Тіл білімі институтында өткен Халықаралық ғылыми-теориялық конференцияның материалдары «Қазақ тіл білімі мен түркітану мәселелері». – Алматы, «Жібек жолы», 2007. 357-360 бб.

Қазақ тілінің қолданбалы саласының даму барысы өмірдің әр қилы тұстарынан байқалады. Солардың бірі – қазақ баспа ісін автоматтандыруға қажетті деген тілдік деректермен қамтамасыз ету тұстары. Сондықтан біз осы шағын мақаламызда, жалпылама сипатта, баспа ісін автоматтандыруға қатысты қандай тілдік деректердің қажеттілігі жайлы сөз етпекпіз.

Баспа ісін автоматтандыру мәселесі авторлық шығарманың түпнұсқасын полиграфиялық түрге келтіруді тек компьютерге жүктеп қана қоймай, сонымен бірге бірқатар лингвистикалық өңдеу әрекеттерімен де айналысатыны белгілі. Қазіргі кезде мұндай әрекеттің төмендегідей кезеңдерін ажырата қарастыруға болады:

1) авторлық кезең;

2) редакциялық кезең;

3) құрылымдау кезеңі;

4) полиграфиялық қайтақұрастыру (воспроизводство) кезеңі;

5) таралымды белгілеу кезеңі немесе мәтінді ұсынылған дана санына сәйкес басып шығару кезеңі;

6) тарату кезеңі.

Аталғандардың ішінен тілдік деректерді аса күшті қажет ететіні екінші кезең, ал тіпті де қажет етпейтін бесінші кезең. Басқа кезеңдерінде лингвистикалық деректерге сүйену орташа ғана дәрежеде болады екен.

Баспа ісін автоматтандыру кезеңдері хронологиялық жағынан бір-біріне тұтасып, бірігіп кетуі де мүмкін. Мысалы, кейбір шығарма авторлары мәтінді өзі теріп, алғашқы түзетулерді өзі жүзеге асыра алады. Сондықтан баспа ісін автоматтандыру кезеңдерінде лингвистикалық деректермен қамтамасыз ету мәселесі әртүрлі сипатта көрініс табады.

Енді кейбір кезеңдердің ерекшеліктеріне қысқаша тоқталайық.

Авторлық кезеңде мәтінді теру үшін авторлар әртүрлі мәтіндік процессорларды қолданады, яғни мәтінді теру мен оған өзгертулер жүргізуде арнайы түрдегі компьютерлік программалар арқылы жүзеге асырады.

Мәтіндік процессорларда тілдің мынадай қызметтері орын алады:

1) автордың қалаған сөзіне синонимдерді іздестіру және қажеттікке қарай, ол сөзді автоматты түрде оның синонимімен алмастыру;

2) қате жазылған сөздерді эталондық орфографиялық сөздіктен іздестіру немесе ол сөздің сөздікте жоқ екендігін айқындау;

3) тұлғалық сәйкестік бойынша эталондық сөздіктегі сөзді мәтіндегі тиісті орнына көшіру.

Авторлық кезеңде қалам иесі қажетті деген мәтін элементтерін тиісті бөліктерге бөліп алуы үшін таңбалар жүйесінің арнайы жазылған формалданған тілін қолдануына мүмкіндігі бар. Егер автор ондай таңбаларды шығарма мәтініне енгізбесе, ол міндетті техникалық редакторлар атқарады. Сонымен бірге бұл кезеңде автордың машиналық ақпарат сақтайтын тетігінде жазылған түсіндірме және энциклопедиялық сөздіктерді автономдық сипатта қолданылуына да мүмкіндігі бар.

Редакциялық кезеңінде баспа ісін автоматтандыру жүйесі мәтінді өңдеуге қатысты корректура жүргізу мен редакциялау қызметтерін жүзеге

асырады. Мәтінді өңдеу жұмысымен қатар редакциялық-баспа қызметі де орындалады.

Баспа ісін автоматтандырудағы корректура мен редакциялау нысаны ретіндегі шығармалар мәтіні мынадай элементтерден тұрады:

– жәй ғана мәтіндер (графемалар, яғни әріптер, морфемалар, сөздер, сөз тіркестері, сөйлемдер, фразадан тыс тұтастық; топтамалар, яғни бөлімдер, тараулар және т.т. және дискурс, яғни бүтін түріндегі мәтін);

– күрделі мәтіндер, яғни математикалық өрнектер мен кестелер;

– шығарма аппаратын құрастырушы мәтіндер (титулдық бет, басылым мәліметтері, сілтемелер және т.б.);

– қызметтік мәтіндер, яғни мәтін элементтерінің таңбалары және полиграфиялық безендіруге қатысты бұйрық элементтері.

Редакциялау кезеңінде баспа ісін автоматтандыру әрекетіне қатысты авторлық мәтінді өңдеудің мынадай екі түрлі амалын ажыратуға болады:

1) мәтіндерді түпнұсқаға сәйкестендіру;

2) мәтіндерді қазіргі кездегі тілдік нормаға сәйкестендіру.

Баспа ісін автоматтандыру ісінде бірінші процесті *корректура* жүргізу деп, ал екіншіні – *редакциялау* деп атайды. Корректура жасау мен редакциялау істерінде де екі түрлі әрекеттерді ажыра білген жөн:

а) бақылау амалы, яғни ауытқуларды табу;

ә) қайтақұрастыру (реконструкция) амалы, яғни бақылау барысында анықталған ауытқуларды түзету.

Бақылау амалының нысаны ретінде барлық мәтін көлемі алынатын болса, ал қайтақұрастыру нысаны ретінде – ауытқулар түрлері орын алатын мәтін бөліктері ғана алынады.

Ауытқу түрлеріне жататындар:

а) бұрмалаулар (түпнұсқадан айырым табатын ауытқулар);

ә) қателер (нормадан ауытқу);

б) қате жіберу мүмкіндіктері (түпнұсқадан да және нормадан да бірдей кететін ауытқулар).

Әрине, редакциялау барысында бақылауды қажет ететін (баспаның, логикалық, техникалық және т.б.) норма мәселесінің барлығы бірдей емес, тек лингвистикалық және психолингвистикалық түрлері ғана орын алады. Бұл топтардың ішінен мынадай нормалық топ түрлерін бөліп қарастыруға болады:

а) тізімдер (мысалы, неше түрлі сөздіктер, қысқартулар тізімі немесе өлшем бірліктерінің тізімі);

ә) үлгілер (мысалы, сөйлемдер мен қайталаным кезіндегі тыныс белгілерді қою үлгісі);

б) құрылымдар немесе модельдер (мысалы, басылымның құрылымы немесе сөйлем моделі);

в) параметрлер (мысалы, бір бетте қатарынан төрт тасымалдың кездеспеуі немесе аннотация көлемінің 600 таңбадан аспау шарты).

Корректурадан немесе редакциялаудан өткен автор мәтіннің сапасын анықтау үшін мынадай шарттардың орындалғаны жөн деп саналады:

– корректура үшін мәтін көшірмесінің бұрмалану дәрежесі мына өрнекпен анықталады:

$$E=e/M.$$

*М*ұндағы ***e*** – мәтінде кездесетін бұрмалану саны, ***M*** – авторлық баспатабақ бойынша есептелген мәтін көлемі;

– редакциялау үшін қателесу дәрежесі:

$$S=r/M.$$

Мұндағы ***r*** – мәтіндегі қате саны.

Машиналық (компьютерлік) корректура жүргізу мен редакциялаудағы кейбір технологиялық ерекшеліктердің ең алғашқысы ретінде мәтінді өңдеудің үш кезеңін бөліп айтуға болады.

Бірінші технологиялық ерекшелікке жататындар:

– корректураның және редакциялаудың алдын алатын әрекеттер.

– интеркорректура (интерредакциялау), яғни компьютер көмегімен орындалатын бақылау мен қайтатуындату әрекеттері;

– корректурадан кейінгі әрекет және редакциялаудан кейінгі әрекет, яғни техникалық құралдар арқылы орындауға келмейтін жағдайдағы корректор мен редактордың амал-әрекеттері.

Екінші технологиялық ерекшелікке корректура мен редакциялаудың әр дәрежедегі автоматтану сипаты жатады:

– компьютерленген корректура мен редакциялау (мәтіндік процессор арқылы компьютер жадына жазылған мәтінге адамның араласуы арқылы ғана жүзеге асатын бақылау мен қайтатуындату әрекеттері);

– автоматтанған корректура мен автоматтанған редакциялау (мәтінді бақылау жұмысы компьютер арқылы жүзеге асады да, ал қайтақұру әрекеті – корректор немесе редактор арқылы);

– автоматтанған корректор мен автоматтанған редакциялау (мәтінді бақылау мен қайтақұру жұмыстары ЭЕМ арқылы орындалады, ал қалыпты жағдаят бұзылса ғана адамның араласуын қажет етеді).

Мәтінді бақылау мен қайтақұрастыруда (қайтатуындатуда) және де корректура мен редакциялау әдістерінің тиімділігін анықтау кезінде мына өрнекті қолдану қажет болады:

$$E=(O_1-O_2)/O_1.$$

Мұндағы ***E*** – мәтінді өңдеудің тиімділігі, O_1 –ауытқулардың саны (мәтінді өңдеуден бұрынғы), O_2 – өңдеуден кейінгі мәтіндегі ауытқулар саны.

Лингвистикалық нормалар, негізінен, қарапайым мәтіндерді тексеруге мүмкіндік жасайды. Мұндай бақылау әдісі әрбір сөзге әртүрлі ақпараттар (морфологиялық, синтаксистік, прагматикалық және т.б.) қоса тіркеліп берілетін сөздіктерге негізделеді. Лингвистикалық әдістердің ең қарапайым түрлері сөздердің кейбір өлшемдері бойынша таңдалып алынған қайсыбір жиынды пайдалануға немесе кейбір сөздерді тіпті де пайдаланбауға негізделеді. Күрделіленген әдістер сөздер арасындағы грамматикалық байланыстардың, тыныс белгілердің, мәтінді абзацтарға бөлудің, қысқартылған сөздер мен сөз тіркестерінің берілулерінің және т.б. дұрыстығын бақылауды жүзеге асырады.

Психо-лингвистикалық нормалардың да барынша кең спектрі қолданыс табуда (*Мучник, 1985)*. Мұндай нормалар белгілі оқырман қауымына арналады және сөз, сөйлем немесе фразадан тыс бірліктердің барынша үлкен «ұзындығын» белгілейді (*тез оқу мүмкіндіктерінің сипаттамалары*).

Құрастыру кезеңінде тілдердің мынадай екі түрі қолданылады:

а) полиграфиялық безендірудің бұйрықтар тілі;

ә) мәтін элементтерін бөліп алуға қажетті таңбалар тілі (язык меток).

Бұл тілдер әртүрлі ұстанымдарға негізделіп жасалады. Мұдай жүйелерде таңбалар мен бұйрықтар саны 20 мен 100 аралығында шектеледі. Көпшілік жағдайда полиграфиялық бұйрықтар келесі құрылымға ие:

а) бұйрық белгісі;

ә) бұйрық түрі;

б) бұйрық параметрі.

Полиграфиялық қайта құрастыру кезеңінде баспа ісін автоматтандырудың лингвистикалық жабдықтау мүмкіндіктерінің ішінен, тек бір ғана функциясы қолданылады. Ол функция сөзді тасымалдау қажеттігі үшін әрекет ететін компьютерлік программа түрінде көрініс табады. Мұндай программалар мәтінді өңдеудегі полиграфиялық безендіру бұйрықтарын жүзеге асыруға арналған автоматты жүйелерде қолданылып жүр.

Сөздерге тасымал жасау программалары баспа ісін автоматтандырудың басқа да жүйелерінде, мысалы, мәтіндік процессорлар мен редакциялық-баспа жүйелерінде қолдануда.

Сөзді тасымалдау алгоритмін мынадай екі үлкен топқа бөліп қарастыруға болады:

– сөздіктер көмегімен, яғни бұл жағдайда компьютер жадында қайсыбір тілдің тасымалданған сөздерінің эталондық нұсқасы сақталады. Мұндай сөздердің көлемі 20-40 мың және бірнеше тілдер түрі (көбінде, оннан аспайтын тілдер) қамтылады;

– сөздіксіз әдіс бойынша, яғни бұл жағдайда тасымалдау әрекеті арнайы ережелердің көмегімен тиісті компьютерлік программалар бойынша жүзеге асады.

Сөздік көмегімен тасымалдау ісі аса нәтижелі болғанымен, ол әдісте әрекет ету мезгілі мен компьютер жадын үнемдеу жағынан көптеп ұтыламыз. Ал сөздіксіз әдіс әрекет уақыты мен компьютер жадына байланысты аса тиімді болғанымен, тасымалдау әрекетінің сапасы өте төмен болады.

Сөздіксіз әдісті мынадай кішігірім топтарға бөліп қарастыруға болады:

– морфемдік әдіс арқылы, яғни тасымалдау белгісін сөздің морфемдік құрамына қарай қою;

– морфемсіз әдіс арқылы, яғни тасымалдау белгісін буындар шекарасына қарай немесе құрамында дауысты және дауыссыз әріптердің екіәріптік тіркестерінің (биграммалардың) болуына қарай қою.

– қисындастыру әдісі. Бұл әдісте тасымалдау белгісі сөзформа құрамындағы префикстердің, суффикстердің және буындардың немесе биграммалардың болуына қарай қойылады.

Баспа ісін автоматтандыруды лингвистикалық деректермен қамтамасыз ету (жабдықтау) жүйесінің құрылымына қысқаша тоқталайық. Мұндай құрылым автономды және жүйелі деп екіге бөлінеді. Лингвистикалық жабдықтаудың автономды түрінде түсіндірме және энциклопедиялық сөздіктер қамтылады. Жүйелі түрдегі лингвистикалық жабдықтауда деректердің тірек базасы, нормалардың жиыны және компьютерлік программалар сақталады.

Деректердің тірек базасының құрамы антонимдік, синонимдік, орфографиялық сөздіктерден, морфологиялық, синтаксистік, семантикалық және прагматикалық ақпараттар сөздіктерінен (жалпытілдік және салалық), эталондық тасымал сөздіктерінен (әртүрлі тілдер үшін), морфемалар тізімінен, өлшем бірліктері қысқартуларының тізімінен, персоналий тізімінен, полиграфиялық бұйрықтар тізімінен, мәтін элементтері таңбаларының тізімінен, библиографиялық сипаттаудың қалыпты элементтерінің тізімінен және арнаулы ақпараттық-іздестіру тілдері рубрикаторларының тізімдерінен тұрады.

Нормалар жиыны іштей былайша бөлінеді: графемдер, морфемдер, фразадан тыс бірліктер, блоктар (топтамалар), дискурстер; жоғары дайындық деңгейіндегі оқырмандар үшін психолингвистикалық нормалар жиыны, төменгі дайындық деңгейіндегілерге – тасымалдау ережелерінің жиыны (әртүрлі тілдер үшін).

Лингвистикалық жабдықтаудың құрастырушылары бөлек және бірге де көрініс табады. Мысалы, лингвистикалық ақпаратты сақтайтын сөздік құрамында орфографиялық және синонимдік сөздіктер бірге берілуі де мүмкін. Осы айтылғанға қарамай, лингвистикалық жабдықтаудың әрбір компонентін жеке бір нысан ретінде қарастыруға болады.

Сонымен, лингвистикалық жабдықтау жүйесіне қойылатын негізгі талап – тек бір ғана лингвистикалық сөздік бойынша әрі синоним сөздерді іздестіруді, әрі корректураны, әрі редакциялауды, әрі сөздерді тасымалдауды жүзеге асыру.

Міне, осы аталған мәселелерді компьютерлік лингвистика тұрғысынан зерттеу қазақ баспа ісін автоматтандыру әрекетін жаңа сипаттағы жоғары деңгейге көтерері сөзсіз.

Әдебиет

Будаар Я. И., Карская С. Д., Партыко З. В. Сравнительный анализ алгоритмов переносов слов русского языка // Научно-техническая информация. Сер. 2. 1985. № 12. С. 20-30.

Мартин Дж. Видеотекст и информационное обслуживание общества. М., 1987. 184 с.

Мучник В. С. Человек и текст: Основы культуры письменной речи. М., 185. 256 с.

Партыко З. В. 1) Методы машинной корректуры и машинного редактирования. М., 1983. 41 с. (Изд. дело: Обзорн. информ. / НИЦ

«Информпечать». Вып.5); 2) Статистика ошибок при корректуре и редактировании текстов. М., 1989. 56 с. (Изд. дело: Обзорн. информ. / НИЦ «Информпечать». Вып.3); Лингвистическое обеспечение автоматизированного издательского процесса / Прикладное языказнание: Учебник, отв. ред. А.С. Герд. СПб: Изд-во С. –Петербург. С. 399-407.

ҚАЗАҚ ЖАЗБА МӘТІНДЕРІНДЕГІ ГРАФЕМА ҚОЛДАНЫСЫНЫҢ СТАТИСТИКАСЫ

ҚР ҰҒА-ның «Хабарлары». Тіл, әдебиет сериясы,
№5(165), қыркүйек–қазан 2007, 28-32-бб.
(Е.Б.Жұбановамен бірге)

Зерттеу нысаны бойынша алынатын таңдама бөліктің ең аз мөлшердегі көлемін анықтау. Фонетика мен фонология түркі тілдер аясында ең көп зерттелген салалар екені белгілі. Мәселен, И.А.Батмановтың, А.Н.Кононовтың, Н.А.Баскаковтың, Қ.Қ.Жұбановтың, І.К.Кеңесбаевтың, Ж.А.Аралбаевтың және т.б. ғалымдардың еңбектерінде түркі тілдерінің фонемалық құрамы жеткілікті түрде зерттелген деуге болады. Түркі тілдерінің салыстырмалы және салыстырмалы-тарихи фонетикасы жайлы ғылыми еңбектерді де көптеп атап кетуге болады. Бірақ бұл зерттеулерде статистикалық әдіс өте сирек қолданыс тапқан. Ал шындығына келсек, қай тіл болмасын, оның фонемалар жиынтығы статистикалық әдіспен талдауға оңай көнетін тілдік бірліктердің қатарына жатады. Статистика-ықтималдықтар пәнінің әдістері арқылы алынған сандық нәтижелер лингвистикалық зерттеулерге сапалы қорытындылар жасауға мүмкіндік тудырады [1; 2; 4, 62-63; 6 және т.б.]. Бірақ кейбір зерттеулер шектеулі көлемдегі мәтін үзінділері бойынша жүргізілгендіктен, олардың нәтижелерінің сенімділік дәрежесі төмен. Сол себепті ондай зерттеулердегі алынған мәліметтерді әр уақытта репрезентативті (өкілетті) деп санай беруге болмайды. Міне, осыған байланысты нысан ретінде алынған мәтін көлемін дұрыс анықтау – аса маңызды мәселеге айналып отыр.

Математикалық статистиканың негізгі міндеттерінің бірі – нысандар жиынтығының қайсыбір қасиеттерін бақылау негізінде бас жиынтықтың (генеральная совокупность) табиғаты жайлы тұжырым жасау. Егер бас жиынтық ретінде қазақ тілінің мәтіндері алынды десек, онда оның көлемі шексіздікке ұмтылатыны дау туғызбайтын жайт.

Математикалық статистика саласында жиынтық бірліктерінің қасиеттерін зерттеу мәселесі үш түрлі тәсіл бойынша жүзеге асады:

1) жиынтық бірліктерінің барлығын бірдей қамту арқылы (жаппай зерттеу);

2) субъективті көзқарас негізінде барлық жиынтық ішінен бірліктердің белгілі бір бөлігін таңдап алу арқылы;

3) барлық жиынтық ішінен белгілі бір бөлікгі (таңдама бөлік немесе таңдама) кездейсоқ түрде бөліп алу арқылы барлық жиынтық (бас жиын) бойынша қорытынды жасалады. Мұндай қорытындыға келу нақты ережелер негізінде ғана жүзеге асады.

Жоғарыда аталған бірінші тәсілді қолдану, яғни жаппай зерттеу жүргізу мәтін көлемінің шексіздігіне байланысты мүмкін бола бермейді. Ал екінші тәсілді, яғни арнайы жолмен іріктеп, жеке көзқарастағы мәтін көлемін таңдап алу әдісін қолданған жағдайда тек шектелген сипаттағы нәтижеге ғана қол жетеді. Мұндай нәтижелер барлық жиынтыққа қатысты статистикалық қорытынды жасауға мүмкіндік тудырмайды.

Сонымен, ең дұрысы, үшінші таңдама әдісті қолдану. Мұндай таңдама бөліктер (таңдамалар) бас жиынның ішкі жиыны деп те аталады және бас жиынның барлық бірліктерінің таңдама ішіне ену ықтималдығы бірдей дәрежеде болады.

Таңдамалар екі түрлі тәсілмен қалыптасады:

а) кездейсоқ сандар кестесі бойынша;

ә) жүйелі түрде іріктеп алу әдісі арқылы.

Мәселен, қазіргі қазақ жазушыларының шығармаларынан үзінділер таңдауда арнайы түзілген «кездейсоқ сандар кестесі» әдісін қолдануға болады. Ол үшін, мысалы, шығарманың бет санын көрсететін нөмірі мен әр беттегі бірінші (немесе басқа) абзац мәтінін таңдамаға енгізу үшін аталған арнайы кестені қолдану керек. Шығарма бетін таңдау арнайы кестедегі сандарға сәйкес келу шартымен жүзеге асады.

Екінші тәсіл бойынша, қажетті зерттеу нысаны ретіндегі материалды іріктеу, яғни әрбір келесі бет саны мен тиісті абзацты таңдап алу жүйелі түрде жүзеге асады. Мұндай іріктеу әдісі, көбінде, газет пен журнал бетіндегі мәтін үзінділерін таңдама бөлікке енгізу кезінде қолданылады. Нысан ретіндегі жеке мақаладағы сөйлемдер осы мақсатпен рет-ретімен нөмірленеді. Содан кейін барып, $K_n=K_{n+1}+L$ формуласы бойынша сөйлемдер реті таңдама бөлік ретінде алынады. Бұл өрнектегі K_n таңдамаға енетін сөйлемнің бастапқы нөмірі ($K_n=1,2,3\ldots$), ал L интервал (арақашықтық) шамасы, әдетте, ол *3* пен *5* сандарына тең алынады. Жүйелі түрде іріктеу кезінде K_n-ның бастапқы шамасы мен интервал мөлшері зерттеушінің өз еркінше алынады.

Статистикалық зерттеудегі таңдама әдісті қолданудың мақсаты – бас жиынтықты сипаттауға мүмкіндік тудыратын деректерді алу. Осыған орай, таңдаманың өкілділігі (репрезентативтілігі) жайында сұрақ (ақиқаттықтың қандай дәрежесімен зерттеу нәтижесін барлық жиынтыққа қатысты сөз етуге болады? – деген сұрақ) туындауы мүмкін..

Математикалық статистика, таңдама бөліктер бойынша жүргізілген зерттеу деректерінің, бас жиынтықтың осындай қасиеттеріне сәйкес келуін бейнелеудің дәлдік дәрежесін анықтауға мүмкіндік береді. Статистикалық талдаудың дәлдігі таңдаманың көлеміне, яғни зерттеуге түсетін бірлік санына айтарлықтай тәуелді болатындықтан, ең алдымен таңдаманың көлемін анықтау мәселесін шешіп алу қажет болады. Аталған мәселе лингвистикалық

материалды статистикалық әдіспен зерттеуде аса маңызды деп саналады [8, 165; 9, 82; 3, 34-35; 7, 76]. Таңдама бөліктің ең кіші көлемдік мөлшерін анықтауға байланысты бірнеше тәсілдер орын алуда [8, 165-166; 3, 33 және т.б.].

Біз өз зерттеуімізде таңдаманың ең кіші көлемін анықтау үшін мынадай өрнекті қолдандық:

$$N = \frac{z_p^2 (1 - f_{гр})}{\delta^2 \cdot f_{гр}},$$

мұндағы Z_p – алдын ала берілген ықтималдықтың сенімділік мәні бойынша алынатын тұрақты сан; $f_{гр}$ – лингвистикалық бірліктің қатынастық жиілігі; δ – қатынастық қате [3, 34-35]. Әдетте, лингвистикалық статистикада ықтималдық мәні ретінде *P=0,95* және *P=0,99* алынады да, ал оған сәкес келетін Z_p-ның кестелік мәні *1,64* және *1,96* немесе *2,58* сандарына тең болып келеді [3, 33]. Ал Z_p-ның кестелік мәндері ықтималдықтар теориясы мен математикалық статистикаға арналған оқулықтардың қосымшасында беріледі.

Қазақ жазба мәтініндегі графемалардың статистикасы. Жазба тілдің сөйлеу тілі сияқты практикалық және теориялық тұрғыдан аса маңызды зерттеу нысаны екені мәлім. Мысалы, ақпаратты автоматты түрде өңдеу мен шартты белгілермен таңбалау (кодтау) кезеңінде сол тілдің графемаларының статистикалық сипаттамаларын білу керек болады. Графемаларды статистикалық әдіспен талдау арқылы алынған нәтижелер баспа ісін жетілдіруде және компьютерді ұтымды пайдалану шараларында да аса қажет. Жазба тіліне тән ерекшеліктер, яғни жазба мәтіннің сандық және сапалық қатынастары қазақ тілінің әртүрлі әдеби жанрларынан орын алады.

Мақалада қазақ тілінің жазба мәтінін әріптік (графемалық) деңгейде статистикалық әдіспен зерттеуге әрекет жасалған. Себебі, әріп – табиғи тілді жазба мүмкіндігі арқылы бейнелейтін ең бір қарапайым түрдегі көрнекі тілдік бірлік болып саналады. Зерттеу нәтижелері қазақ тілін фонологиялық және фонетикалық деңгейде тілдің морфемдік құрылымын талдау мен тілдік бірліктерді синтагматикалық тұрғыда қарастыру жағдайларында аса маңызды. Әріптердің, әріп тіркестерінің жалпы және шептік орналасу мүмкіндіктерінің статистикасын анықтауда А.Х.Джубановтың (қазір – А.Қ.Жұбанов) «Задача получения на ЭВМ частотных списков лингвистических единиц» [5] атты мақаласында келтірілген алгоритмдік сызбасы компьютерлік программа жазуға негіз болды. Сонымен бірге, аталған автордың «К вопросу о графемной статистике казахского текста» [6] атты еңбегіндегі әріптердің статистикасы жайлы мәліметтер де осы мақаланы жазуға ықпал етті.

Жиілік сөздіктерді құрастыруда қазақ тілінің үш түрлі стильдер материалдары зерттеу нысаны ретінде алынды. Олар:

1) М.О.Әуезовтің «Абай жолы» романы мәтіні (*51290* сөзқолданыс немесе *280812* әріп):

2) математика пәніне арналған мектеп оқулықтарының мәтіні (*19467* сөзқолданыс немесе *130388* әріп);

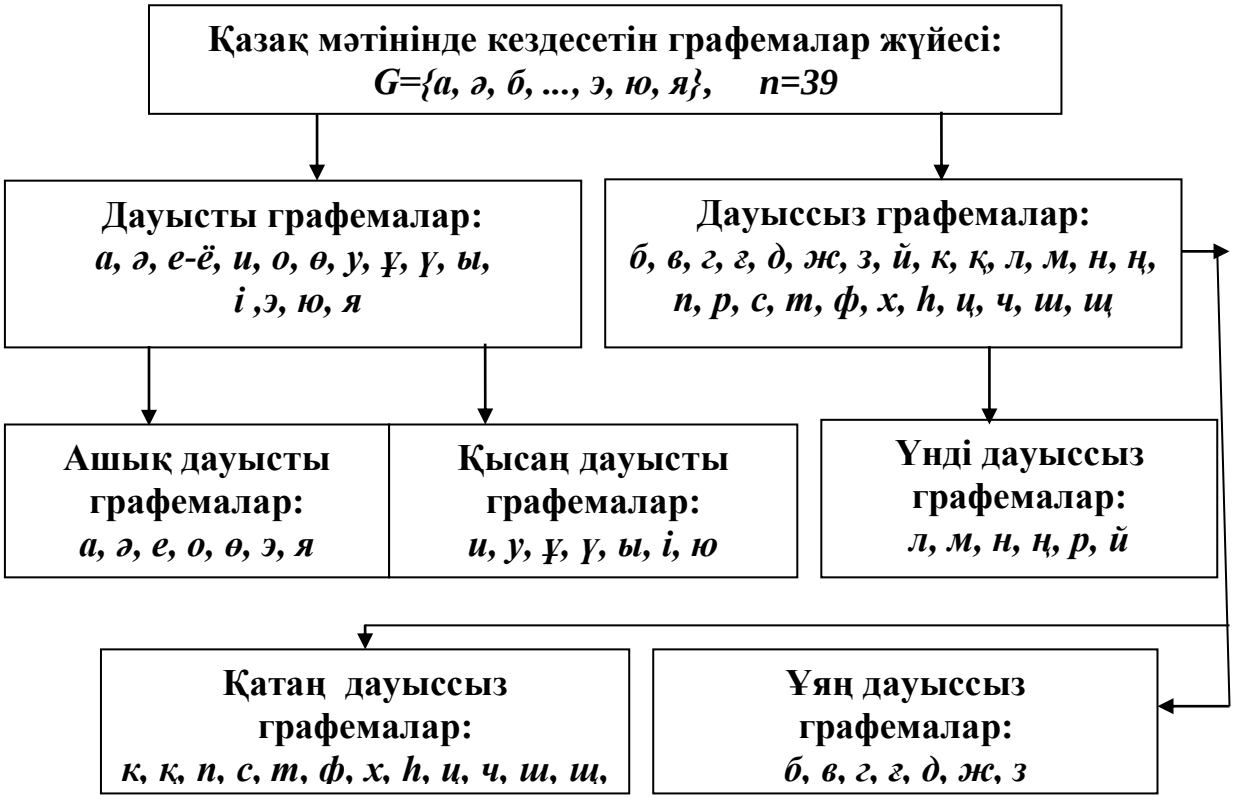
3) қазақ тілінің екі томдық түсіндірме сөздігінің мәтіні (*17585* сөзқолданыс немесе *116317* әріп)

Әрі қарайғы пайымдауымызда, бұл стильдерді қысқаша түрде – *роман, математика, сөздік* деп атауды жөн көрдік.

Тәжірибе түрінде алынған стильдер мәтіндерінің тең көлемді болмауы, олардың толық түрдегі шығармалар болуымен түсіндіріледі. Сондықтан мәтіндер ішіндегі тілдік бірліктердің (әріптердің) сандық сипаттары жайлы мәлімет алу үшін олардың қатынастық жиіліктері (немесе пайыздық салмақтары) пайымдалды.

Мәтін әріптерін статистикалық әдіспен талдаудан бұрын олардың графемалық құрамын анықтап алған жөн. Мысалы, егер қазақ тілі әліпбиіндегі барлық әріптерді (графемаларды) *G* жиыны деп белгілесек, жиын элементтерінің саны *n*-ге тең болады. Бұл жиынға (*G*) мәтіндегі сөз бен сөздің арасын ажыратып тұратын «бос орын» белгісін және қос сөздер арасындағы «дефисті», «тыныс белгілерін» есепке алмағанда, әліпбидегі жіңішкелік белгі (*ь*) мен қатаңдық белгіні (*ъ*) бір графема деп есептесек, онда *n=39* деуге болады. Яғни *G={а, ә, б, ..., э, ю, я}* және жиын элементтерінің саны *n=39*.

Зерттеу барысында мәтінде кездесетін әріптер жиынтығы топ-тобымен ірілі-ұсақты бөліктерге бөлініп (шартты түрде – дауысты, дауыссыз), төмендегі сызба-топтама түрінде көрініс тапты.



Сызба-топтама

Болашақ зерттеу жоспарымызда сызбада көрсетілген графемалар (әріптер) мен олардың тіркестерінің мәтінде қолданылу мүмкіншіліктерінің статистикасы, әртүрлі стильге қатысты мәтіндер бойынша және олардың сөздегі шептік орналасу тәртібіне қарай да (сөз басында, сөз ішінде, сөз соңында) зерделенеді. Ал қазіргі ұсынып отырған мақаламызда біз тек қазақ әріптерінің белгілі көлемдегі мәтін бойын қамту мүмкіндіктерінің пайызына байланысты ғана жайттардың статистикасына тоқталамыз.

Төменде ***1-кестеде*** қазақ тілінің үш стилі бойынша компьютер көмегімен түзілген қазақ әріптерінің (графемаларының) жиілік сөздігінен үзінді келтірілді. Қысқаша пайымдағанда, аталған кестеде ең жиі кездесетін графемалар олардың жиіліктерінің кему тәртібімен орналасқан. Әрі қарайғы пайымдауымызда графемалардың үш түрлі стиль бойынша түзілген жиілік сөздіктеріндегі алғашқы он орынына қатысты статистикалық мәліметтер, яғни әрбір әріптің және олардың жиынтықтарының мәтінді қамту мүмкіндіктерінің пайыздық салмақтары сөз болады. Кестедегі мәліметтерге сүйенсек, қазақ әліпбиіндегі *39* графеманың тек 10-ы ғана әр стиль бойынша алынған мәтіндердің *64-66* пайызын қамтитынын байқаймыз. Ал көркем әдебиет (роман) стилі мен ғылыми-техникалық (математика) стильде ең жиі қолданатын төрт графема ғана (*а, е-ё, ы, н*) барлық мәтіннің *35-36* пайызын қамтиды екен. Бұл жерде әмбебап дауыстылар қатарына жататын *4*-ші “*і*” дауысты графема аталған стильдерде 5-ші және 6-шы орындарға (іргелес орындарға) ие болып тұр.

1-кестедегі сөздік мәтіні (қазақ тілінің екітомдық түсіндірме сөздігі) бойынша алынған әріптердің жиілік сөздігіндегі ең жиі қолданыстағы әріптердің алатын орны мен статистикасы жоғарыда қысқаша қарастырылған екі стиль бойынша алынған мәліметтерден белгілі дәрежеде айырым табады. Әсіресе, сөздіктегі тұйық раймен берілген етістіктердің әсерінен «*у*» әрпінің 3-орынға ие болуы және «*қ*» әрпінен басталатын қазақ сөздерінің сөздіктегі басымдылығы бірден көзге түседі. Ал жиілік сөздіктегі алғашқы ең жиі кездесетін «*у*» мен «*қ*»-дан басқа 8 графема орналасу орындарымен ғана айырым тапқанымен, мәтінді қамтудағы пайыздық салмағы жағынан айтарлықтай айырым таппайды.

Әрине, қазақ графемаларының толық жиілік сөздігіндегі мәліметтер жеке пайымдауды қажет етеді. Ал біз бұл мақалада жиілік сөздік үзіндісіндегі мәліметтермен бірге, қазақ тілінің дауысты және дауыссыз графемалар топтарының қысқаша статистикасын да беруді жөн көріп отырмыз (2-кесте және 3-кестені қара).

Сөздің басында да, соңында да және ішінде де қолданыла беретін әріптерді – «сөз ішіндегі» қолданыс деп шартты түрде атауды ұйғардық. Міне, осындай дауысты және дауыссыз дыбыстарды таңбалайтын графемалардың статистикасын анықтауда, олардың қатынастық жиіліктерінің стильдік (жанрлық) айырым-белгі ретінде жүрмейтіндігін 2-ші және 3-ші кестедегі деректерден аңғаруға болады. Әрине, сөздік мәтініндегі реестр сөздердің дыбыс құрамының статистикасы қиындасқан сөздер тізбегіндегі статистикадан аздап болса да айырым табады.

Жеке графемалардың жиілігін талдай келе, кейбір дауыстылардың және дауыссыздардың басқа графемаларға қарағанда жиі қолданатынын байқау қиын емес. Оған мысал ретінде «әмбебап» дауыстылар деп аталатын *(а, е-ё, ы, і)* графемаларды айтуға болар еді. Мұндай графемалардың мәтін ішінде қолданылуының қосынды нәтижесі: романда – *34,8%;* математикада – *32,0%;* сөздікте – *29,0%.* Егер біз қарастырған мәтіндердегі барлық дауысты графемалардың қолданылу жиілігінің пайыздық қосындысы – *43,14%; 43,19%* және *44,60%* екенін ескерсек, әмбебап аталатын *а, е-ё, ы, і* төрт дауысты графемалардың қолданылу дәрежесін өте жоғары деп санауға әбден болады.

1-кесте

Стиль-дердегі әріптің алатын орны	Әріптердің мәтінді қамту пайызы (%)								
	Роман			Математика			Сөздік		
	Әріп аты	Әріп-тің %	Жиын-тық %	Әріп аты	Әріп-тің %	Жиын-тық %	Әріп аты	Әріп-тің %	Жиын-тық %
1	а	13,2	13,2	а	11,0	11,0	а	13,0	13,0
2	е-ё	8,1	21,3	е- ё	8,3	19,3	т	7,1	20,1
3	ы	7,5	28,8	н	8,2	27,5	у	7,0	27,1
4	н	7,1	35,9	ы	7,4	34,9	е-ё	6,3	33,4
5	і	6,0	41,9	д	5,4	40,3	ы	6,1	39,5
6	т	5,0	46,9	і	5,3	45,6	р	6,0	45,5
7	р	4,8	51,7	р	5,3	50,9	л	6,0	51,5
8	д	4,6	56,3	т	5,2	56,1	с	4,5	56
9	л	4,2	60,5	л	5,1	61,2	қ	4,1	60,1
10	с	4,1	64,6	с	4,4%	65,6	н	4,1	64,2
Қосындысы:			65%			66%			64%

2-кесте

Дауыссыз графемалардың топтары	Дауыссыз графемалардың мәтінді қамту пайызы (%)		
	Роман	Математика	Сөздік
Үнді	22,60%	24,60%	21,20%
Қатаң	19,95%	17,71%	23,20%
Ұяң	14,31%	14,50%	11,00%
Қосындысы:	56,86%	56,81%	55,40%

3-кесте

Дауысты графемалардың топтары	Дауысты графемалардың мәтінді қамту пайызы (%)		
	Роман	Математика	Сөздік
Ашық	25,93%	25,27%	24,40%
Қысаң	17,21%	17,92%	20,20%
Қосындысы:	43,14%	43,19%	44,60%

Дауыссыз графемалардың статистикасына байланысты айтатынымыз, қазақтың жазба және сөйлеу мәтіні сипатына үнді және жабысыңқы-шұғыл графемалардың тән болуы. Ал шұғыл (үзілмелі) графемалар көркем әдебиет мәтінінде *23,5%* тең, математика мәтінінде – *22,19* және сөздік мәтінінде *22,2* пайызға тең екен және бұндай пайыздық шамалар барлық графемалардың төрттен біріне ғана жуық деуге болады.

Кейбір *в, х, ч, h, щ, ц, э, ю, я* – графемалар және *ь-ъ* белгілері басқа тілдерден (орыс, араб-иран) енген кірме сөздерде ғана кездесуіне байланысты, олардың пайыздық салмағы бір пайызға да жетпейтін дәрежеде қолданылған.

1, 2, 3-кестелер бойынша басқа да мәліметтерді сөз етуге болатыны белгілі, бірақ олар келесі мақаламыздың нысаны болмақ.

Әдебиет

1. *Байтанаева Д.А.* Информационные характеристики казахского текста: Автореф. дисс. ... канд. филол. наук. Алма-Ата, 1985. 17 с.

2. *Бектаев К.Б.* Статистико-инженерные методы в тюркологии // СОПИЯЛ. Чимкент, 1980. С. 7-10.

3. *Бектаев К.Б.* Статистико-информационная типология тюркского текста. Алма-Ата: Наука, 1978. 183 с.

4. *Головин Б.Н.* Язык и статистика. М.: Просвещение, 1971. 191 с.

5. *Джубанов А.Х.* Задача получения на ЭВМ частотных списков лингвистических единиц. В кн.: Статистика казахского текста. Алма-Ата, 1973. С. 263-299.

6. *Джубанов А.Х.* К вопросу о графемной статистике казахского текста. В кн.: Вопросы казахской фонетики и фонологии. Алма-Ата: Наука, 1979. С. 49-52.

Исенгельдина А.А. Факторы, определяющие относительную частотность фонем. В кн.: Статистика казахского текста. Алма-Ата, 1973. С. 659-662.

7. *Нелюбин Л.Л.* Перевод и прикладная лингвистика. М.: Высшая школа, 1983. 207 с.

8. *Пиотровский Р.Г.* Моделирование фонологических систем и методы их сравнения. М.Л.: Наука, 1966. 299 с.

9. *Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.С.* Математическая лингвистика. М.: Высшая школа, 1977. 384 с.

10. *Фрумкина Р.М.* Статистические методы изучения лексики. М.: Наука, 1964. 116 с.

III БӨЛІМ

ҚАЗАҚ ТІЛІНІҢ «ТІЛ – ҚАЗЫНА» АТТЫ ҰЛТТЫҚ КОМПЬЮТЕРЛІК ҚОРЫ

***

НАЦИОНАЛЬНАЯ КОМПЬЮТЕРНАЯ БАЗА «ТІЛ – ҚАЗЫНА» КАЗАХСКОГО ЯЗЫКА

ҚАЗАҚ ТІЛІНІҢ «ТІЛ – ҚАЗЫНА» АТТЫ ҰЛТТЫҚ КОМПЬЮТЕРЛІК ҚОРЫНЫҢ МҮМКІНДІКТЕРІ ЖАЙЫНДА

ҚР ҰҒА «Хабарлары», филология сериясы.
Алматы, 2008, №2, - 19-23 бб.

Соңғы жылдары ғылыми-техникалық прогрестің дамуын жеделдетуге компьютерлік технологиямен бірге барлық ақпараттық индустрия да өз әсерін тигізе бастады. Қоғам өмірінде ақпарат ағымының таралуы табиғи тілде жүзеге асатындықтан, тіл білімінің қызметі күннен күнге артуда. Тілдің жаңа компьютерлік формасы пайда болып жатыр. Тілдің мұндай формасында өзіне тән заңдылықтар да жоқ емес. Мысалы, оның өз «диалектісі» мен кәсіби жаргоны, тұрақты грамматикалық құрылымы, өзіне ғана тән лексикасы мен орфографиясы бар деуге болады.

Қоғамдағы ақпарат алмасу жүйесінің ең алғашқы жетістігі ретінде баспа станогін ойлап шығару саналса, екінші ақпараттық революция деп электронды-есептеуіш мәшиненің (компьютердің) күнделікті тұрмысымызға араласуын айтуға болады. Әрине, тіл білімі тұрғысынан бұл құбылыс тіл табиғатына жаңаша көзқарасты қажет етеді. Мәселен, кітап бастыру, кітаптарды тарату және сонымен бірге сауаттылық деңгейін арттыру әрекеттері ұлттық және әдеби тілдің қалыптасуында айтарлықтай қозғаушы күш болғаны мәлім. Сондықтан қазіргі заманда компьютерге бағдарлама

жаза білу (программирование) екінші сауаттылық деп аталып жүр. Бұл құбылыстарды ғылыми тұрғыда ұғыну яғни табиғи тіл мен жасанды тіл арасындағы қатынасты ғылыми жолмен зерттеу қолданбалы қажеттіктен кешігіңкіреп дамып келеді. Алайда, «адам-мәшине» аралық қарым-қатынаста және табиғи тілдегі мәліметтерді автоматты түрде өңдеу тәжірибесінде қажетті деген тілдік деректермен қамтамасыз ету қазіргі заманғы ақпараттық жүйелердің ажыратылмас бөлігіне айналып отыр. Осыған байланысты тіл білімінің алдында тың мәселелер тұрғаны белгілі. Дәлірек айтсақ, олар: туындаған жаңа мәселелерді арнайы зерттеу, оларды табиғи тілдің белгілі заңдылықтары мен категорияларына сәйкестендіру; «адам-мәшине» аралық қарым-қатынас пен табиғи тілдегі деректерді автоматты түрде өңдеуде табиғи тілдің заңдылықтары мен категорияларын жинақтап қорыту; аталған жүйелерді қажетті сөздіктер мен грамматика деректерімен қамтамасыз ету. Бұл мәселелердің шешімінің табылуы теориялық немесе нақты қолданбалы мәні болумен қатар әлеуметтік және халықшаруашылық мәндері де өте маңызды.

Егеменді еліміздің ақпарат аясындағы ғылыми-техникалық дамуының болашақ жобаларында табиғи тілдегі, соның ішінде қазақ тіліндегі деректерді өңдей алатын бағдарламалармен жабдықталған дербес компьютерлердің жаңа буыны өмірге келетіні сөзсіз. Бұл компьютерлер жады өте көлемді болумен қатар «білім банкі» деп аталатын анықтамалық және сөйленіс түріндегі ақпараттық құралдармен жабдықталуы қажет.

Қарым-қатынастың компьютерлік формасына электрондық құралдар арқылы шығарылатын құжаттардың, мақалалардың, кітаптардың, газет-журналдардың, кесте мен сызба мәтіндерін пайдалануға болады. Сол сияқты жобалаудың, бір тілден екінші тілге аударма жасаудың, анықтамалық қызмет етудің, сырқат адамдарға диагноз қоюдың және т.б. бұқаралық сипаттағы қарым-қатынастың автоматтанған формасына көшуді жазудың жаңаша бір түріне ауысу жағдаятымен салыстыруға болар еді. Жазу формалары адамдар қарым-қатынасына ғаламдық сипат бергені мәлім. Бұл ұрпақтар жалғастығын нығайта түсті деуге болады. Ал жазудың компьютерлік формалары ақпарат тарату ісіне ғарыштық деңгейдегі электрондық жылдамдықты қамтамасыз етті.

Академик В.М.Глушков «Основы безбумажной информатики» атты кітабының алғы сөзінде былай деп жазған болатын: «Большинство выполненных к настоящему времени прогнозов сходятся в том, что к началу следующего столетия в технически развитых странах основная масса информации будет храниться в безбумажном виде – в памяти ЭВМ. Тем самым человек, который в начале XXI в. не будет уметь пользоватся этой информацией, уподобится человеку начала XX в., не умевшему ни читать, ни писать» [2].

Ақпараттық индустрияның ЭЕМ-ге қатысты даму қарқынын сипаттайтын кейбір мәліметтерге тоқталайық.

Кітап бастыруда дүние жүзі бойынша 500 жылдық даму кезеңінде орта есеппен бір жыл ішінде бір адамды екі кітаппен қамтамасыз етсе, ақпарат

индустриясының 40 жылдық дамуы барысында әлемнің әрбір тұрғынына бір дербес компьютер шығарылады екен. Қоғамға қажетті ақпараттың ең көп бөлігі компьютерлік ортада пайда болып, ұзақ мерзімге сақталады әрі оны тез арада тауып алуға мүмкіндік туады. Сол себепті тілдік деректерді автоматты түрде жедел өңдейтін ЭЕМ-ді құрастырып, табиғи тілде жүзеге асыру мәселесі заман сұранысы болып отыр.

Есептеуіш лингвистиканың бұл мәселелермен айналысқанына 50 жылға жуық уақыт өтті. Осының негізінде көптеген орыс және басқа да шет тілдерге қатысты күрделі мәселелердің шешімі табылды. Олардың қатарында ғылыми-техникалық әдебиеттер мен өндірістік деңгейдегі құжат мәтіндері бойынша мәшинелік аударма жасау және т.б. жатқызуға болады. Әрине, бұл саладағы жетістіктерді әлі де жетілдіре түсу керек.

Тілшілер қауымы зерттеу жұмыстарын осы қарқынмен және осындай ұйымастыру жағдайында жалғастыра беретін болса, ақпараттық аядағы ғылыми-техникалық прогрестің даму процесі айтарлықтай бәсеңдеуі мүмкін. Сол себепті зерттеу барысында кездескен кедергілерді жоюды ғылыми тұрғыда қарастырып, есептеуіш лингвистиканың барлық жетістіктерін біртұтас бағдарламаға біріктіріп, жаңа ізденістерді сол бағдарламаға сүйене жүргізу керек. Дәстүрлі және қолданбалы лингвистика арасындағы алшақтықты жоюға әрекет етіп, қазақ тіліне қатысты академиялық деңгейдегі еңбектердің негізінде мәтіндерді автоматты түрде өңдеу жүйесін құрастыру қажет-ақ.

Бұлшық ет қызметі мен ойлау жұмысын автоматтандырудың жедел қарқынмен өсуі әлеуметтік-саяси және экономикалық жүйелердегі адам факторының маңыздылығын арттыра түсті. «Адам-мәшине» жүйесі бойынша қарым-қатынаста ойлау әрекетінің түрлері, тіл, әлеуметтік және мәдени деңгей, білімдарлық пен хабардар болу мәселелері, өз тарапынан, білімді жинақтау құралдарын дамытуды қажет етеді. Қоғамның технологиялық қызметінде лингвистиканың маңызы артты. Мәселен, қоғамдық ғылымдар, жаратылыстану мен техникалық ғылымдардың өзара әрекеттестігінде тіл білімінің белсенділігі күшейді. Нәтижесінде қолданбалы тіл білімінің *формалды-математикалық* және *есептеуіш лингвистика* деп аталатын ғылым салалары өмірге келді. Осы бағыттар арқылы қазақ тіл біліміне формалдау, алгоритмдеу, құрылымдау (конструирование) идеялары мен жаңа әдістер еніп, қолданыс тауып келеді. Осындай ұғымдық аппарат арқылы тіл білімінің жетістіктері шекаралас жатқан ғылым салаларына да таралды. Қазіргі синтаксис және семантика, лексикология және лексикография, сөйлеу тілі арқылы мәтін теориясы сияқты лингвистикалық ұғымдар ақпараттық бағыттарда қолданыс тауып келеді:

- бағдарлама жазу тілдері мен деректер қоры теориясында;
- ғылыми-техникалық әдебиеттер мен құжаттарды автоматты өңдеу тәжірибесінде;
- ойлау қабілеті бар роботтарды құрастыруда және т.б. сияқты.

Ғылым салаларының мұндай өзара әрекеттестігінің болашақта да белсенді түрде дами беретіні сөзсіз. Оған нақты әлемдік адам-мәшинелік

қарым-қатынас және құжаттарды автоматты өңдеу қажеттілігі ғана емес, сонымен бірге тіл білімі жинақтаған зор білім қоры негіз болды. Осы білім қорын есептеуіш жүйелерде жүзеге асыратын болсақ, қазіргі компьютерлер шын мәнінде адамдармен зияткерлікпен әрекеттесу мүмкіндігіне ие болар еді.

Қазақ тілінің әлеуметтік-саяси, мәдени және технологиялық қызметінің, құжаттар мен әдебиеттерді, тілді қолданудың жаңа формаларына сәйкес *көркем әдебиет*, *ғылыми-көпшілік әдебиет* және *баспа тіл әдебиеті* деген дәстүрлі түсінік те қазіргі кезде жаңа мәнге ие болды. Бүгінгі заманда әдеби тіл ұғымына көркем әдебиет, публицистика, сөйлеу тілі стильдерімен бірге ресми, ғылыми және ғылыми-техникалық жазба тілі, техникалық құжаттар тілі ұғымдарын қоса есептеу қажет болып отыр. Әдеби тілдің құрамында мұндай құраушылардың болатындығын елемеу, тілдік бірліктерді нормалау (кодтау, таңбалау) нәтижелерінің тілдік қарым-қатынаста таралмай қалуына және көпмағыналыққа әкеп соғады. Кейбір лексикографиялық еңбектерге ғылыми-техникалық стильдердегі көптеген арнайы сөздер мен сөз тіркестерінің енбей қалуына байланысты өндірістік тілдік қарым-қатынастағы мәтіндердің мазмұны дұрыс ашылмай қалып жүр.

«Адам-мәшине» қарым-қатынас жүйесінде қолданылатын компьютерлік сөздіктер мен бағдарламалар үшін негізгі дереккөз ретінде академиялық сөздіктер мен грамматикалар алынуы қажет деп санаймыз. Алайда аталған дереккөздер мәдени және танымдық құндылықтарына қарамастан, тілдік фактілерді қамтуы жағынан да, дәлме-дәл баяндалу талабына қатысты да жоғарыда аталған талаптарға сай келе бермейді. Сондықтан қазақ тіліндегі деректерді автоматты түрде өңдеу жүйелерін құрастыруда академиялық басылымдарға ғана сүйенбей, іскери және ғылыми-техникалық жазба материалдары бойынша жүргізілген жеке зерттеулер нәтижелерін де негіз етіп алу қажет. Қазақ тілі аясындағы жүргізілетін теориялық зерттеулер мен қолданбалы ізденістерді қазіргі заманғы практикалық қолданыс талаптарына сәйкес келетіндей бірыңғай әдіснамалық негізде жүргізіп, өзара жақындастырған жөн. Бұл тілдік деректердің сөздіктер мен грамматикаларда толық әрі дәл қамтылуын ғана көздемей, сонымен бірге ол еңбектердің қай формада берілуіне де ерекше көңіл бөлуді қажет етеді. Полиграфиялық пішіндегі дәстүрлі сөздіктер мен грамматикаларды электрондық түрге келтіру жұмысы қиын да күрделі екенін ескерсек, олардың электрондық формада көрініс табуы компьютер жадында автоматты өңдеудің қажеттігінен туындап отыр.

Тіл білімінің қолданбалы бағыты оның теориясын дамытып, тереңдете түсуде ерекше маңызды болумен қатар, әртүрлі ғылым салаларының өзара әрекеттесіп, бір-бірін толықтырып отыр. Оған мысал ретінде символдық тізбектерді формалды өзгерту теорияларын да алуға болады. Лингвистикада ондай теориялар *туындату грамматикасы* (порождающая грамматика) және *трансформациялық грамматика* деп аталып жүр. Бұл теориялар ең алдымен дискретті математиканың бастауында пайда болып, есептеуіш мәшинелерді қолдануда, компьютерлік бағдарлама жазуда кең қолданыс тапты. Оның

жекелеген ерекшеліктері табиғи тіл деректерін өңдеуде де нәтижелі болуда.

Лингвистика мен ақпараттанудың (информатиканың) өзара әрекеттесуі екі салаға да оң әсерін тигізеді. Тіл зерттеу тәжірибесіндегі үлкен жобаларды жүзеге асыру, көптомдық сөздіктер мен грамматикаларды өмірге келтіру талпыныстары осы салалардың теориялық жетістіктері деуге де болады. Бірақ осындай зерттеу жұмыстарымен айналысатын ғылыми ұжымның мүмкіндігі арасында кейбір қайшылықтарда жоқ емес. Мәселен, 20-40 жыл бойы атқарылатын үлкен жобаларды жүзеге асыруда көбінде, ғалымдар тобы тұрақты болмай, ауысып отырады. Міне, осындай жағдайда, алғашқы тұжырымдаманы сол қалпында сақтау мүмкін бола бермейді. Тіл білімінің қолданбалы саласында құрастырылатын автоматты сөздіктер мен грамматикалар күрделілігі жағынан дәстүрлі сөздіктерден кем түспейтіне қарамастан, оларды автоматтанған әдістермен аз уақыт ішінде дайындап шығаруға болады.

Алғашында теориялық және қолданбалы лингвистикалар арасындағы, жасанды түрде әдетке айналған сапалық айырмашылықтың бүгінгі күні сандық пішінге айналуы, автоматтандыру әдістерін филология саласына енгізуде бірқатар кедергі болуы мүмкін. Тіл білімі осындай алшақтықты жоюға әрекет етуі қажет. Ол үшін ғылым мен техниканың есептеуіш және тіл білімінің қолданбалы саласындағы қол жеткен жетістіктеріне сүйене отырып, лингвистикалық зерттеулерді кешенді түрде автоматтандыратын және тиісті ақпаратпен қамтамасыз ететін жүйелер құрылғаны жөн.

Кеңес үкіметі кезінде орыс тілінің мәшинелік қорын жасау жолдары зерттеліп, қажетті жоба жасалған болатын. Бұл жобаны іске асыру үшін лингвистикалық зерттеулерді кешенді автоматтандыру ісі, орыс тілінің барлық лексикалық байлығын сыртқы магниттік таспада жинастыру, лингвистикалық алгоритмдер мен бағдарламалар қорын құру, орыс тіліндегі мәтіндерді автоматты түрде талдау мен жинақтау жүйесінің толық аяқталған қорын құру және тіл білімі бойынша бірнеше ақпаратты-анықтамалық жүйе құру жұмыстары алдын ала жоспарланған болатын. Әрине, мақаламызда орыс тілінің мәшинелік қоры бойынша жасалған жобаны талдауды мақсат етіп отырған жоқпыз. Бұл орайда ол жұмыстардың бәрінің жүзеге аса қоймағаны, кейбіреулерінің әліде зерттеліп жатқанын ғана айтпақпыз. Біздің міндетіміз – жобаның тәжірибелеріне сүйене отырып, қазақ тіліне үйлесімді жақтарын анықтау, қазақ тіл білімін автоматтандыруға қатысты қолданбалы саласының жаңа бағытын дамыту.

Осыған байланысты төмендегілерді атауға болады:

- **Қазақ тілінің «Бас сөзтізбесі».** Мұндай сөздер тізімінде қазақ тілі бойынша жасалған картотекаларда, сөздіктерде, энциклопедиялар мен басқа да дереккөздердегі сөздер мен тұрақты сөз тіркестерінің тізбесі қамтылады;

- **Қазақ тілінің терминологиялық қоры.** Бұл қорға салалық ақпараттық жүйелер мен терминологиялық стандарттарда тіркелген барлық лексиканың жиынтығы енеді. Сонымен бірге А.Байтұрсынұлы атындағы Тіл білімі институтында «Ұлттық идея – Қазақстанның даму негізі» бағдарламасына қатысты 2007 жылдан бастап орындалып жатқан “Қазақ

тілінің терминологиялық қорының компьютерлік базасын жасау және оны терминдік әлеуеті жоғары дереккөздермен толықтыру” тақырыбының ғылыми нәтижелері де осы қордың еншісіне жатады.

- **Академиялық сөздіктер мен грамматикалар қоры.** Бұнда автоматты режімде қолданыс табатын қазақ тілінің академиялық сөздіктері және академиялық грамматикалардың деректері беріледі. Мәселен, А.Байтұрсынұлы атындағы Тіл білімі институтында «Тәуелсіз Қазақстан жағдайында қазақ тілінің ұлттық және мемлекеттік тіл ретінде негізгі даму үрдістері мен бағыттары» атты негізгі ғылыми бағдарлама бойынша 2006 жылдан бері зерттеліп жатқан «Қазақ тілінің компьютерлік базасы: лексикалық қордың ашық жүйесі» атты тақырып бойынша алынатын нәтижелерді осы қордың үлесіне жатқызуға болады. Оған: а) «Абай тілі сөздігі»; ә) біртомдық «Қазақ тілінің сөздігі»; б) Салалық терминдердің түсіндірме сөздіктері; в) А.Байтұрсынұлы атындағы Тіл білімі институтында құрастырылып жатқан «Қазақ әдеби тілінің сөздігінің» (15 томдық түсіндірме сөздіктер) материалдары жатады.

- **Қазақ тілінің иллюстративті-мәтіндік қоры.** Бұл қорда қазақ тіліне қатысты негізгі мәтіндік топтамалар орналасады. Солардың негізгісі ретінде қазіргі кезде А.Байтұрсынұлы атындағы Тіл білімі институтында жасалып жатқан қазақ тілінің «Тіл – қазына» атты ұлттық компьютерлік базасының картотекалық қорын атауға болады. Қағаз-кәртішкі түріндегі бұл мәнмәтіндік қордың көлемі 5 млн. жуық. Әрине, бұл мәнмәтіндік қор тиісті сапалық және сандық тұрғыдағы сұрыптаулардан өткеннен кейін оның көлемі біршама азайды. Қазіргі кезде осы қордың «А» әрпінен «С» әрпіне дейінгі сөзтізбелерге қатысты мәнмәтіндер «Тіл – қазына» атты ұлттық компьютерлік базаға енгізілді, басқа әріптерге қатысты реестр сөздердің мәнмәтіндері де компьютерлік қордан тиісті орнын алады.

Жоғарыда аталған қорларды жаңа технологиялық негізге ауыстыру нәтижесінде оларды қарқынды түрде жаңа дерекқорлармен толықтыруға болады. Осының нәтижесінде сөздіктер мен анықтағыштар жедел түрде баспадан қайта шығарылып, көпшіліктің пайдалануына беріледі. Компьютерлік графиканың мүмкіндіктерін қолдану нәтижесінде диалектологиялық атластарды құрастырып, баспадан шығару және тарихи жазба ескерткіштерді өңдеу ісін автоматтандыруға да жағдай жасалады. Осының нәтижесінде түбегейлі жаңа сөздік пайда болып, мәтіні қазақ тілінің «Тіл – қазына» ұлттық компьютерлік қорына енеді. Сөйтіп, әрі қарайғы зерттеулерді дамытуға, құрылымдық қолданыстарын кеңейтуге жол ашылады.

Лексикографиялық жұмыс нәтижелерін жақсартудың міндетті шарты – «лексикографиялық құралдарды» алдын ала дайындау болып табылады. Оған қазақ лексикасының морфемдік сөздігін құрастыру, оны семантикалық жағынан топтастыру, тіркесімдік, қолданылу жиілігін анықтау және т.б. жатады. Аталған жұмыстарды автоматтандыру мәселесі қазақ тілінің «Тіл − қазына» ұлттық компьютерлік қорының үлесіне жатады. Ұлттық қор материалдары зерттеу жұмыстарында ғана пайдаланылумен қатар қазақ тілін

оқыту, баспа ісінде және т.б. да кең түрде қолданыс табатыны сөзсіз.

Қазақ тілінің «Тіл – қазына» ұлттық компьютерлік қоры тілдің қолданбалы жақтарын толық қамту нәтижесінде қазақ тілінің анықтамалық қызметінің тиімділігін арттырып, тілдің теориялық жақтарын зерттеуде аса құнды материал бола алады. Себебі аталған компьютерлік қор қазақ тілінің бүгінгі күйімен қатар тарихи грамматика мен диалектологияны, психолингвистика мен әлеуметтік лингвистика мәселелерін зерттеуге, тілдердің жалпы даму заңдылықтарын ашуға да көмектесе алады.

Қазақ тілінің «Тіл – қазына» ұлттық компьютерлік қоры лексикография саласының дамуына, жазушылар тілін, дара тілдік шығармашылықты зерттеуде де ерекше маңызды. Сонымен бірге бұқаралық сөйлеу тіліне, баспа тіліне, радио мен теледидар тілдерін жүйелі түрде зерттеуге де мүмкіндік береді.

«Тіл – қазына» атты ұлттық компьютерлік қорды жүзеге асыру лингвистикалық зерттеулерді автоматтандыру, деректерді өңдеудің өндірістік жүйелерін тілдік мәліметтермен жабдықтауда жаңа бетбұрыс деуге болады.

Әрине, «Тіл – қазына» жобасында ең алдымен компьютерлік ортада қызмет ететін жалғамалы қазақ тілінің өзіне тән әдіс-тәсілдерін қарастырып алу қажет болады. Зерттеудің формалды және математикалық әдістерін, олардың нәтижелерінің берілу формасын барынша толық зерттеп алған жөн. Себебі формалды көрініс пен математикалық дәлдікте берілген нәтижелер ғана компьютерлік «Тіл – қазына» қорының игілігі бола алады және сол арқылы басқа зерттеушілер мен тұтынушылардың нәтижелерді дұрыс түсінуіне мүмкіндік береді. Сөздіктер базасындағы лексикографиялық деректердің формалды көрінісі (берілісі) арқылы ғана әртүрлі сөздіктерді салыстырып, өзара біріктіруге болады.

Автоматты аударма жасау жүйелеріндегі формалды әдістер қазақ тілі мен орыс тілі арасындағы өзара әрекеттестікте аса маңызды. Мәселен, қазақша-орысша, орысша-қазақша автоматтанған сөздіктер құрастыруда, автоматты аударма жасауда және т.б. тіларалық зерттеулерде аса құнды. Ол үшін шет тілдері бойынша жүзеге асқан автоматтанған аудармадағы және автоматты лексикография салаларындағы қол жеткен нәтижелерді қазақ тілінің лексикография тәжірибесіне жедел түрде енгізу қажет.

Қазақ тілінің «Тіл – қазына» атты ұлттық компьютерлік қорын құрастыру мен дамыту ісіне жас маман-зерттеушілердің атсалысқаны жөн болар еді. Ол үшін Қазақстан жоғарғы оқу орындарындағы филология факультеті жанында құрылымдық, қолданбалы және математикалық лингвистика бойынша мамандықтар мен кафедралар ашылуы қажет-ақ. Ондай жоғары оқу орындарын жаңа технологиялық құралдармен қамтамасыз етіп, оқу жоспары мен ғылыми-зерттеу жұмыстарының тақырыптарын жаңартып, қазақ тілінің «Тіл – қазына» атты ұлттық компьютерлік қорын дамыту үшін мамандар дайындау заман талабы. Сонымен бірге қазақ тілі мамандарына есептеуіш тіл білімі мен лингвистикалық зерттеулерді автоматтандыратын техникалық құралдармен әрекет ету әдістерін үйрететін

мамандарды даярлау ісінде академиялық ғылыми институттар мен жоғарғы оқу орындары тығыз байланыста болуы қажет-ақ.

Лингвистикалық зерттеулер мен ізденістерді кешенді автоматтандыру жүйесі ретіндегі қазақ тілінің «Тіл – қазына» атты ұлттық компьютерлік қорын толыққанды жүзеге асыру, шындығында да, қазақ тілінің байлығын көрсетеді, оны *тілдің қазына қоймасы* деуге де болатындай.

ӘДЕБИЕТ

1. *Андрющенко В.М.* Концепция и архитектура Машинного фонда русского языка. – М.: Наука, 1989. – 200 с.

2. *Глушков В.М.* Основы безбумажной информатики. – М.: Наука, 1982.

3. *Жұбанов А.Қ.* Тіліміздің компьютерлік қорын жасау керек //Ана тілі. – 2003. – 13 қараша (№ 45).

4. *Жұбанов А.Қ.* Қазақ тілінің компьютерлік қорын (базасын) құрастыру жайлы // Жұбанов А.Қ. Компьютерлік лингвистикаға кіріспе: Оқу құралы. – Алматы: Қазақ университеті, 2007. – 204 б. (189-199 бб.).

ЖАҢА АҚПАРАТТЫҚ ТЕХНОЛОГИЯҒА КӨШУ ЖАҒДАЙЫНДАҒЫ ҚАЗАҚ ЛЕКСИКОГРАФИЯСЫ

ҚР ҰҒА «Хабарлары», филология сериясы.
– Алматы, 2008, №5, - 14-18 бб.

Қазақ тілінің лексикография саласындағы сөздік құрастыру ісінде жаңа ақпараттық технологияға көшу лексикография мамандарын көптеген қосымша жұмыстардан босатады деп білеміз. Жаңа ақпараттық технологияға сөздік картотекасын автоматтандыруды, автоматты сөздіктер базасын құрастыруды, сөздік мақаламен компьютер экраны арқылы әрекет етуді және т.б. мәселелерді жатқызуға болады.

Қазақ тілін зерттеу тәжірбесінде лексикографиялық жұмыстарды ЭЕМ көмегімен автоматтандыру мәселесі 1973 жылдан ғана бастама алды. Солай бола тұра, қолданбалы лингвистиканың статистикалық және компьютерлік салалары А.Байтұрсынұлы атындағы Тіл білімі институтында айтарлықтай орын алуда [1; 2]. Мәселен, 1993 жылы «Қолданбалы лингвистика» бөлімінің қызметкерлері М.Әуезовтің 20 томдық толық шығармалар жинағын компьютер жадына түсіріп, зерттеуші қауымына неше түрлі жиілік сөздіктер ұсынған болатын [3]. Міне, осындай жиілік сөздіктердің негізінде ұлы жазушының тіл байлығы мен стильдік ерекшелігі жайында бірнеше ғылыми зерттеулер жүргізіліп, диссертациялар қорғалды. Келесі тәжірбие ретінде істелген ғылыми жұмыс – көп жылдар бойы жинақталып-теріліп, қазақ тілінің бірден-бір «тіл байлығы» деуге болатындай, институттың «сөз қоржынында» сақталған қазақ тілінің картотекалық қорын сан мен сапа жағынан сұрыптап, компьютер жадына түсіріп, автоматты түрде пайдалану

ісі. Аталған жұмыстың қажеттігі тек бүгінгі таңға ғана емес, оның болашағының маңыздылығында. Кәртішкі-мәнмәтін түріндегі қордың жалпы көлемі, орташа есеппен алғанда, 5 млн. шамасындай. Әрине, осындай үлкен көлемдегі мәнмәтін қорын автоматты жүйеге айналдырмайынша, оны тілші қауымының ойдағыдай пайдалану мүмкіндігі, сол күйінде сақтап тұруы немесе үнемі жаңартылып отыруы оңайлыққа түспейді.

Міне, осындай себептерден өткен ғасырдың 90-шы жылдарынан бастап қолға алынған қазақ тілінің «Қазына» (қазіргі «Тіл – қазына») атты автоматтандырылған картотекалық қордың маңызы айтарлықтай зор. 1991-1993 жылдар аралығында «Қолданбалы лингвистика» бөлімінің қызметкерлері «Қазына» қорына 900 мың кәртішкі-мәнмәтін ішінен 70 мыңдайын сұрыптап алып, оларды компьютер жадына енгізген болатын. Қазіргі кезде бастапқы әріптері «А» әрпінен «С» әрпіне дейінгі сөздерге қатысты мәнмәтіндер «Тіл – қазына» компьютерлік базаға енгізіліп, тұтынушы-лексикограф үшін қажетті деген мәліметті іздеп табу жұмыстары жүзеге асырылды. Ал «Т» әрпінен «Я» әрпіне дейінгі кәртішкі-мәнмәтіндерді база енгізу әрекеті өз жалғасын табуда.

«Тіл – қазына» атты қазақ тілінің кәртішкі-мәнмәтіндік компьютерлік базаның негізгі мүмкіндіктерін қысқаша сөз ететін болсақ, олар:

- қажетті деген сөзге немесе сөзформаға қатысты кәртішкі-мәнмәтіндерді лезде тауып алып, олармен компьютердің экраны арқылы немесе қағаз бетіне басып шығарып та танысу мүмкіндігі;

- әрбір әріпке байланысты реестрлік сөздік түзілгендіктен, қажет деген сөздер мен оларға қатысты кәртішкі-мәнмәтіндерді лезде іздеп тауып алу мүмкіндігі. Мұндай әрекетті кәртішкі-мәнмәтін ішіндегі сөзқолданыстар бойынша да жүзеге асыру алдын ала қарастырылған;

- қажетті деген сөзге немесе сөзформаға (олардың бастапқы, ортаңғы, соңғы бөліктеріне) байланысты кәртішкі-мәнмәтіндерді авторына, шығармасына, баспа аттары мен шыққан жылдарына қатысты іздеп табу мүмкіндігі;

- «Тіл – қазына» компьютерлік базадағы кәртішкі-мәнмәтіндерді өңдеуге, өшіруге немесе жаңадан қосуға да мол мүмкіндік бар.

Қазақ тілінің «Тіл – қазына» атты компьютерлік базасы картотекалық қордың барлық міндетін атқарумен бірге, ол енді сөздіктер мен мәтіндерді бейнелеу қорына да айналып отыр. А.Байтұрсынұлы атындағы Тіл білімі институтының іргелі (2006-2008жж.) және қолданбалы (2008-2009жж.) зерттеу бағдарламалары бойынша «Компьютерлік лингвистика» ғылыми тобының атқарып жатқан ғылыми жұмыстарының нәтижелері осының айғағы. Қазақ лексикасының түсіндірме сөздіктері негізіндегі компьютерлік базасын іске қосу мақсатына сай зерттеу жұмысының нысаны Тіл білімі институтында құрастырылып, 1968 жылы жарық көрген «Абай тілінің сөздігі» мәтіні (6 мың сөз, 1 мыңдай сөз тіркес), 1999 жылы баспадан шыққан «Қазақ тілінің сөздігі» (50 мың сөз) және жаңадан құрастырылып жатқан 15 томдық «Қазақ әдеби тілінің сөздігінің» алғашқы төрт томының мәтіні. Сол сияқты «Ұлттық идея – Қазақстанның даму негізі» бағдарламасы бойынша:

«Қазақ тілінің терминологиялық қорының компьютерлік базасын жасау және оны терминдік әлеуеті жоғары дереккөздермен толықтыру» атты ғылыми-зерттеу жұмысының негізгі нысаны салалық екітілді (қазақша-орысша, орысша-қазақша) терминологиялық сөздіктер мен қазақ тіліндегі салалық терминологиялық түсіндіме сөздіктер материалдары.

Мұндай компьютерлік қор жасаудың негізгі мақсаты – зерттеуші-тілшінің жұмыс орнын толығымен автоматтандыру. Осының арқасында зерттеушінің уақытын барынша үнемдеу, яғни қажетті тілдік ақпаратты іздеу жұмысы мен тілдік бірлікті реттеу және т.б. жеңілденетіні сөзсіз.

Сонымен бірге тілдік процестерді автоматтандыру мәселесі тоқырап қалмай, үнемі жетілдіру жағынан өзінің жалғасын тауып тұруы қажет. Егер ана тілімізге күн сайын жаңа сөздер, жаңа қолданыстар мен жаңа терминдердің еніп отыратынын ескерсек, мұндай зерттеу – күн тәртібінен түспейтін әрі ұтымды, әрі заман сұранысына сай зерттеу деп білеміз.

Автоматты сөздік базасын құрастыру мәселесі қазақ лексикографиясы үшін толығымен шешілген жағдайда, сөздікті екі түрлі тұрпатта – кітап түрінде және сыртқы ақпарат сақтау таспаларына жазып алынған электрондық пішінде сақтауға мүмкіндік туады. Нәтижесінде сөздіктердің қолдану аясы кеңейеді және лексикография саласының қызметі де ұлғая түседі.

Жаңа ақпараттық технологияны қазақ лексикография саласына қолдану үшін ең алдымен автоматты лексикографияның дамуы мен кейбір теориялық мәселелерін айқындап алу қажет деп білеміз. Осы саланың белгілі маманы В.Г.Гак лексикография қызметін былайша ажыратады [4, 14-15]:

- тіларалық қатынас;
- тілді үйретуге қатысты;
- ана тілді сипаттау мен нормалауға байланысты;
- тілді ғылыми тұрғыда зерттеуге байланысты.

Аталған қызметтердің басымдылығына қарап, В.Г.Гак сөздік ісін кезеңкезеңге бөліп қарастырады: сөздікке дейінгі кезең (Еуропада – XV ғ. дейін); ерте сөздік кезеңі (Еуропада – XV-XVII ғғ.) және дамыған лексикография кезеңі (Еуропада – XVIII ғ. бастап).

Сөздікке дейінгі кезең глоссалармен, вокабуляриялармен, ономастикондармен және сол сияқты тіл мәдениеті саласының кейбір мағынасы күңгірт сөздерге түсініктеме беру қызметін атқаратын шығармаларымен сипатталады. Ерте сөздік кезеңі лексикондармен және ұлтаралық қатынаспен және кең мағынадағы оқыту қызметін атқаратын екітілді сөздіктер арқылы түсінік алады. Ал дамыған лексикография кезеңі аталған қызметтердің бәріне де қатысы бар сөздіктердің әртүрлі типтерімен, әсіресе түсіндірме сөздіктер түрімен көрініс табады.

Осы мәселені XX ғ. мен XXI ғ. басы аралығына қатысты сөз еткенде кем дегенде екі қызметін тағы да атап кетуге болады. Олар: ақпараттық жүйелерді лингвистикалық қамсыздандыру және терминологиялық қызмет ету (XX ғ. екінші жартысына тән); кең мағынадағы адам-компьютераралық

қатынас (редакциялық-баспа қызметін автоматтандыру ісін қоса есептегенде).

Лексикография қызметін кезең-кезеңге бөлуде назар аударатын тағы бір жайт, ол, шамамен алғанда, сөздік көлемінің әр кезең сайын бір деңгейге жоғарылауының жетекші қызметінің де ауысуына әкелуі. Дегенмен, мақала көлемінің 0,3-0,5 миллион шамасы полиграфиялық пішіндегі сөздіктер үшін ең жоғарғы шек болып саналатыны белгілі. Себебі, ондай көлемдегі сөздіктер көпшілікқолды болуға өте көлемді, аса күрделі, бағасы жоғары және тез арада оны толықтыру мен баспадан қайта шығаруға қолайлы емес. Ең алдымен, олар әртүрлі тілдік қызметті атқаратын анықтағыштық құрал рөлін атқарудан гөрі, сол дәуірдің тілі мен лесикография саласының ескерткіштері деуге көбірек келетін сияқты.

Лексикография саласының негізгі технологиялық тәсілі – сөздік мақаланы координаттық үзінділеу (фрагменттеу) мен сілтемелік аппарат қолдану. Аталған тәсіл арқылы есептеуіш лексикографияда жалпылама сипаттағы компьютерлік бағдарламалау әдісі де қолданыс тауып келеді. Ал сөздіктің математикалық құрылымы деген ұғымға келетін болсақ, оны белгіленген тораптар (сөздік мақала үзінділері) мен доғаларға (сілтемелер) бағышталған граф деуге болады. Мәселен, оған сәйкес келетін есептеуіш құрылымдар – тізімдер және бағдарлама арқылы жүзеге асатын тізім құру мен іздеу механизмі. Бұл әдісті қазақ лексикографиясында жүйелі түрде қолдану әртүрлі лексикографиялық мәселелерді шешуге мүмкіндік туғызар еді. Мысалы, бір ғана алгоритм арқылы қазақ тілінің автоматты сөздігінен (АС) сөзтіркестердің барлық типін (терминологиялық, фразеологиялық, тұрақты сөз тіркестерін) іздеп табуға мүмкіндік туады. Ал сөздік мақаланы ықшамдау әрекетіне байланысты сөзжасамдық, дескрипторлық және идеографиялық (тақырыптық) сөздіктерге бір ғана әдісті қолдануға мүмкіндік туады.

Қазақ тіліне қатысты қажетті ақпаратты компьютердің жадында сақтау мен оны іздеп табу бір ғана құрылымдық алгоритм бойынша жиілік сөздіктер үшін де, сөзнұсқағыштар мен конкорданстар үшін де бірдей дәрежеде қолданылатын болса, ондай құрылым морфемдік сөздіктер үшін де өзгеріссіз қалады. Сол сияқты, деректер құрылымы мен көптілдік АС-ны компьютерлік бағдарламамен қамтамасыз ету тәсілдерін ассоциативті (байланысты) сөздіктер үшін де қолдануға болады.

Сонымен, қазақ тілінің автоматтанған лексикография саласын төмендегідей біркелкі сипаттағы типологиялық топтарға бөліп қарастыруға болады:

1. Сөзнұсқағыштар, жиілік сөздіктер, конкорданстар, морфемдік сөздіктер;

2. Орфографиялық және орфоэпиялық сөздіктер, «бірге немесе бөлек» жазылатын сөздер сөздіктері;

3. Семантикалық сөздіктер, сөзжасамдық сөздіктер, дескрипторлық сөздіктер түрлері;

4. Барлық аударма және ассоциативтік сөздіктер;

5. Барлық түсіндірме, энциклопедиялық, этимологиялық, көптілді терминологиялық сөздіктер және түсіндірме-комбинатарлық сөздік;

6. Барлық ортологиялық сөздіктер, паронимдер, синонимдер және антонимдердің сөздіктері.

Осы топтардың әрбіреуі сөздік мақаланың бірыңғай қорытындыланған құрылыммен және деректерді компьютер жадында орналастыру мен оларды іздестіру стратегияларымен сипатталады.

Сөздік базасында мәліметтерді сақтау мен тауып алу және олардың алгоритмдік сызбасына қатысты мәселелер жоғарыда аталған алғашқы бес типологиялық топ үшін жеткілікті түрде зерттелді деуге болады.

Әлем бойынша ең нәтижелі зерттеулер терминологиялық лексикографияға қатысты жүргізілгені мәлім. TEAM, LEXIS, EURODICAUTOM, TERMIUM сияқты терминологиялық банктердің әрбіреуінде көлемі бір миллионнан асатын терминдік бірліктер сипатталған [5, 185-193]. Есептеуіш техника мен ядролық физика салаларына қатысты орыс, ағылшын, француз, неміс тілдері бойынша 60 мың сөздік мақала көптілді МУЛЬТИЛЕКС атты автоматты сөздіктің (1985 жылғы версиясының) жадында сақталған. УНИЛЕКС атты автоматтанған лексикалық жүйенің құрамына осы АС-ның кеңейтілген ТЕРМИН түрі сөздік базасы ретінде енген [6]. Ал бұл жүйенің мәтіндік базасы автоматты конкорданс пен жиілік сөздіктерді туындататын бағдарламалар пакетінен тұрады.

LEDA атты автоматтанған лексикалық жүйе де жоғарыда аталған жүйелермен ұқсас құрылымда деуге болады. Жүйенің мәтіндік банкінде 20 млн. сөзқолданыс сақталған, ал оның 13 млн-ы – жазба тіл корпусы, ауызекі тіл корпусы және газет тілі корпусы; 7 млн-ы – әр баспалардан жинақталып алынған күйінде сыртқы магниттік құрылғыларда сақталуда. Қазіргі кезде мәтіндік базаның көлемі 50 млн. сөзқолданысқа дейін өсті. Аталған жүйеде неміс тілі деректерінің әртүрлі сөздіктері бойынша қалыптасқан сөздік базасы да орын алған. Оның жалпы сипаттағы сөздік мақаласы бас сөздің (реестрдің) транскрипциясын, лемманы, сөз табын, диагностикалық қосымшаны, сөзжасам мен синтаксистік қоршауы жайлы және т.б. мәліметтерді өз құрамында сақтайды.

Сөздік түрлерінің көп болуы тілдік қызметтерді лексикографиялық материалдармен жабдықтау мәселесін еш уақытта күн тәртібінен ығыстыра алмайды. Сондықтан ең дұрыс шешім – интегралдық әдіске үстемдік беру, яғни деректердің лексикографиялық базасына (ДЛБ) артықшылық беру. Себебі оның артықшылығы, біріншіден, бар сөздік қорды және оған қатысты барлық әрі қажетті, әрі қоғамдық деңгейде аса құнды ақпаратты өз жадына сыйғыза алуы және екіншіден, тұтынушының қалаған ақпаратын еш өзгеріссіз қажетті формада тауып беру мүмкіндігінің болуы.

Қазақ тіліндегі лексикографиялық жұмыстарды автоматтандыру ісінде Ю.Н.Карауловтың «Лингвистическое конструирование и тезаурус литературного языка» [7] атты кітабының маңызы аса зор деуге болады. Аталған еңбекте тұңғыш рет лексикографиялық параметрлерді жүйелеу

әрекеті жүзеге асырылған. Сөздіктердің микро- және макроқұрылымдарын параметрлеу лексикографиялық ақпараттарды және оған сәйкес келетін деректердің сөздік базасын жобалаумен бірге әртүрлі сыртқы құрылымдағы сөздіктерді де салыстыруға мүмкіндік туады. Бейнелі сөзбен айтқанда, көптілді сөздіктерде сөздік мақалының өзегі деп «түсінік берудің семантикалық бірлігі» алынса, ал аударма сөздік типтерінің өзегі ретінде «аударма бірлігі» алынады деуге болады. Мұндай тұжырымдар түсіндірме және аударма сөздіктердің атқаратын қызметінен туындап тұр: егер түсіндірме сөздіктердің атқаратын мәселесі, негізінен алғанда, сөздік бірлігінің мағынасын ашып беру болса, аударма сөздіктердің ондай мәселесі – сөзқолданыстың контекстуалдық жағдайына тәуелді болып келетін, аударма баламасын табу. Мәселен, аударма (терминологиялық) сөздіктердің түсіндірме сөздіктерден айырмашылығы – реестр сөздердің 80 пайызы тұрақты терминологиялық тіркестерден тұруы және сөздік мақала құрылымының артықшылығы – контекстегі сөзжасамдық пен тіркесімдікке қатысты аударма сәйкестігін бейнелейтін ұялы құрылым. Сондықтан қазақ тіліндегі деректердің сөздік базасын пайдалану кезінде бірсөзді және көпсөзді контекстуалды сауалға қатысты мақаланы толық қамтудың орнына, тұтынушы қалауына сәйкес келетін қажетті бөлігін ғана компьютерлік бағдарлама көмегімен айқындап, жауап түрінде ұсынуға болады. Терминологиялық сөздіктерде бұл мәселе стандарттардағы немесе сапалы аударма бар екітілді автоматты конкорданстардағы ақпаратты қамту (қатыстыру) арқылы жүзеге асады. Түсіндірме сөздіктерде мұндай мүмкіндік сауал контексін талдау арқылы жүзеге асады. Мысалы, ондай мүмкіндіктер бойынша сөздіктегі түсіндірме соңында берілетін типтік иллюстрациялардың негізінде тек сауалды ғана қанағаттандыратын түсініктеме бөлігін жекелеп алуға болады.

Сөздік ақпаратының фрагменттері (үзінді бөліктері) жинақталып қортындыланған және формалданған түрде болулары қажет. Мысалы, морфологиялық ақпарат қисынды межеленген түрде болса, олар арқылы сөздер мен сөзтіркестерді анализдеу мен синтездеу бағдарламалары бойынша фонетикалық (орфографиялық) алмасулар мен ерекшеліктерді және кейбір жүйесіздікті ескеретін қажетті деген қопарманы (флексияны) таңдауға болады. Сонымен бірге, тұтынушы үшін дәстүрлі сөздіктерде қабылдаған ақпаратты да (сөз таптары, типтік қосымшалар) қалыптастыруға мүмкіндік туады.

Автоматты сөздіктерде формалданбаған мағынаның ауызша дефинициясымен бірге мағынаны формалды сипаттау әдісі де қолданылғаны жөн болады. Ал әдеттегі сөздіктердегі сөзқолданыстың синтаксистік және контекстуалдық шарттарға бағыттайтын иллюстрациялардың орнына лексикалық және синтаксистік тіркесімге қатысты әрі айқын, әрі формалданған мәліметтер қолданылуы қажет. Мұндай мәліметтер екіжақты рөлді атқара алады: біріншіден, олар автоматты анализдеу мен синтездеу процедурасында қолданылса, екіншіден, деректердің мәтіндік базасындағы

сөзқолданыстардың типтік иллюстрациясын таңдау әрекетінде де орын алуы керек.

Сөйтіп, сөздік типінің өзі де өзгеріске ұшырайды. Сөздік қатал формалды болуы қажет, яғни оның түп бейнесі қазіргі кездегі автоматты аударма жүйесіндегі немесе жасанды иттеллект жүйесіндегі сөздіктерге ұқсас болуы шарт.

Сөздік мақаланың ұялы, идеографиялық немесе әліпби тәртібімен орналасуына байланысты дәстүрлі мәселелер автоматты лексикографияда өз мәнін жоғалта бастағаны мәлім. Сол сияқты, сөзтіркестер мен фразеологизмдердің бір немесе бірнеше сөздік мақалада қайталанып берілуі және лексика-семантикалық варианттарға қатысты сөздік мақалаларда өзара сілтемелердің кездесуі де АС үшін аса маңызды емес деуге болады. Деректер базасында сілтемелердің еркін және автоматты түрде қалыптасуы мен оларды оқу әрекетінің электрондық жылдамдығы сөздік «жүгін» азайтумен бірге қажетті ақпаратты іздеп табу мезгілін үнемдеу мәселесін де барынша жеңілдетеді. Тұтынушы сөздіктің ішкі құрылымына айтарлықтай назар аудармай-ақ, соңғы нәтижесімен ғана (қажетті ақпаратпен ғана) танысуына мүмкіндігі бар.

Дәстүрлі лексикография саласына аян болып келетін сөздік мақала құрамы бойынша типтеу, лексемалардың грамматикалық және лексика-семантикалық типіне қарай лексикографиялық ақпаратты түр-түрге бөлу автоматты лексикографияда айырықша сипатқа ие болуда. Аталған мәселе компьютердің жадында сақталатын ақпаратқа ғана қатысты болумен бірге бағдарлама жасаудың алгоритмдік және логикалық жақтарын зерделеуде де аса маңызды.

Автоматты лексикографияда сөзтізбені іріктеу, түсіндірмелік пен энциклопедиялық, жалпытілдік пен оқу, анықтағыш пен нормативтік сөздіктердің типтерін ажырату да айтарлықтай маңызды емес. Автоматты сөздіктерде орын алатын мәліметтер лингвистикалық тұрғыда түсініктеме беруге мүмкіндігі бар немесе нақты аударма түрінде баламасы бар, немесе тіпті басқаша бір жолмен түсініктеме беруге болатындай болуы керек. Әрине, лексиканың әртүрлі типі, әртүрлі түсіндірілуі, әр әдіспен иллюстрациялардың берілуі және т.б. автоматты сөздік ішінде алдын ала келісілген тәртіппен таңбалануы керек. Осы таңбалардың арнайы рөлі бағдарлама жазуда ескеріледі және олар нақты тұтынушының мұқтаждығын қанағаттандыруға да бағышталған. Ал тұтынушы қатарында: ғалым-лексикограф, оқушы-студенттер, көркем әдебиет оқырмандары, ғылыми-көпшілікке арналған әдебиет пен ғылыми-техникалық әдебиет оқырмандары, белгілі бір пәнді жақсы, орташа, нашар меңгерген аудармашылар, сол сияқты шет тілін жақсы немесе нашар білетін маман-оқырман және т.б. болуы мүмкін екендігін ескеру қажет [8, 24].

Сонымен, қазақ тілінің автоматты лексикография саласының өмірге келуі дәстүрлі сөздіктердің кемшін тұстарын толықтыра алады деуге болады. Жоғарыда сөз болғандай, дәстүрлі нормативті сөздіктерде кейбір сөздердің түсіп қалу себебін екі түрлі мағынада түсіндіруге болады: біріншісі – әдеби

нормадан тыс болуынан, екіншісі – сөздердің толық қамтылмауынан. Осы сияқты қысқарған сөздерді, сөзжасамдық қосымшаларды, жалқы есімдерді, олардың ішінде географиялық атауларды, мекеме аттары мен өнеркәсіп бұйымдары аттарын және т.б. сөздерді сөздікке қосу-қоспау мәселесі де өз шешімін табуы керек. Қазақ тілінің автоматты сөздіктерінде тұтынушының сауалдарына жауап беру әлеуеті мүмкіндігінше молынан қамтылуы қажет, себебі басқаша жағдайда, ол сөздік өзін-өзі ақтай алмауы да мүмкін.

Қазақ тілінің автоматты лексикография саласының жаңа лексиканы ескеруде де, «масыл» сөздерден «құтылу» мәселесінде де қазіргі қоғам мен өнеркәсіп қызметінде кең қолданыс тауып жүрген баспа, теледидар, радио немесе интернет желісі арқылы қадағалап тұруына және соның нәтижесінде толық шешімге келу мүмкіндігі мол. Сондықтан қазақ тілінің лексикография саласындағы автоматтандыру тәсілі сөздік ісінің технологиялық барысын жақсарта түсуге барынша атсалысады. Тек автоматтандыру әдісі арқылы ғана қазақ тілінің сөздік құрастырушылары арасындағы еңбекті бөлісу ісі дұрыс жүзеге асады. Нәтижесінде әрбір лексикограф сөздердің грамматикалық, лексика-семантикалық және тақырыптық жағының біркелкі сипаттағы белгілі топтарына ғана жауап беретін болады. Сөздікті дайындау процесін тек осылай ұйымдастыру арқылы ғана құрастыру мезгілін барынша қысқартуға болады деп білеміз.

Қазақ тілінің автоматты лексикографиялық базалары тілші-лексикографтарды немесе кез келген тұтынушыны қажетті деген ақпаратпен жабдықтау және тілдік материалдарды іздеп табу әрекетін барынша жеңілдететіні сөзсіз. Сөздік мақала құрылымын шаблондау (бір үлгіге келтіру) және үлгіленген дайындаманы шағын көлемдегі сөздіктерге реттілікпен ұсыну сөздіктің жүйелілігі мен нұсқаулыққа сай келуін қамтамасыз етеді. Ал нұсқаулықтар ақпараттық-есептеуіш жүйе арқылы әр уақытта бақылауда болатын формалданған жобалық тапсырмаға айналады. Міне, осының нәтижесінде сөздік құрастыру барысы үлкен ғимараттарды, ірі техникалық объектілер мен жүйелерді автоматты жобалау процесіне ұқсас жүзеге асады деуге әбден болады.

Әлемдік өндірісті интеграциялау кезеңінде әртүрлі түсініктерді бірізге түсіру ісі де аса қажет болуда. Сондықтан егемен Қазақстанның халықаралық деңгейге шығуына және бірқатар жетекші елдермен қазіргі таңдағы ғылым мен технологияның әртүрлі салаларында ғылыми және өндірістік ақпараттармен алмасудың едәуір ұлғаюына байланысты неғұрлым кең ауқымды қамтитын көптілді сөздіктер шығару қажеттілігі де туындауда.

Міне, осы аталған жайттар қазақ тілінің компьютерлік қорын жасауды жаңа технологияға көшу негізінде жүзеге асыру бүгінгі күн тәртібіне қойылатын ең маңызды мәселелердің бірі деуге болады.

ӘДЕБИЕТ

1. Қазақ тексінің статистикасы. I шығарылуы: Алматы: Ғылым, 1973. 731 б.;

2. Қазақ тексінің статистикасы. II шығарылуы: («Лингвистикалық жұмыстарды автоматтандыру және тіл статистикасы» тобы қызметкерлерінің еңбектері). Алматы: Ғылым, 1990. 208 б.

3. *Бектаев Қ.Б., Жұбанов А.Қ., Мырзабеков С.* М.О.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері. -Алматы-Түркістан, 1995. 346 б.

4. *Гак В.Г.* О некоторых закономерностях развития лексикографии (Учебная и общая лексикография в историческом аспекте) // Актуальные проблемы учебной лексикографии. М.: Рус. яз., 1977.

5. *Волкова И.Н.* Стандартизация научно-технической терминологии. М.: Изд-во стандартов, 1984

6. *Андрющенко В.М.* Автоматизированная лексикографическая система UNILEX (основные проектные решения) // Вычислительная лингвистика. М.: Изд-во МГУ, 1982.

7. *Караулов Ю.Н.* Лингвистическое конструирование и тезаурус литературного языка. М.: Наука, 1981.

8. *Шайкевич А.Я.* Проблемы терминологической лексикографии // Обзор. информ. Сер. 1. Теория и практика науч.-техн. пер. М.: ВЦП, 1983. Вып. 8.

АВТОМАТТЫ (МАШИНАЛЫҚ) АУДАРМА

Алматы қаласының Тілдерді дамыту басқармасы
Алматы қаласы әкімдерінің «Тіл» оқу-әдістемелік
Орталығы аударматану (ғылыми-практикалық
көмекші құрал). Алматы, 2008. 83-111-бб.

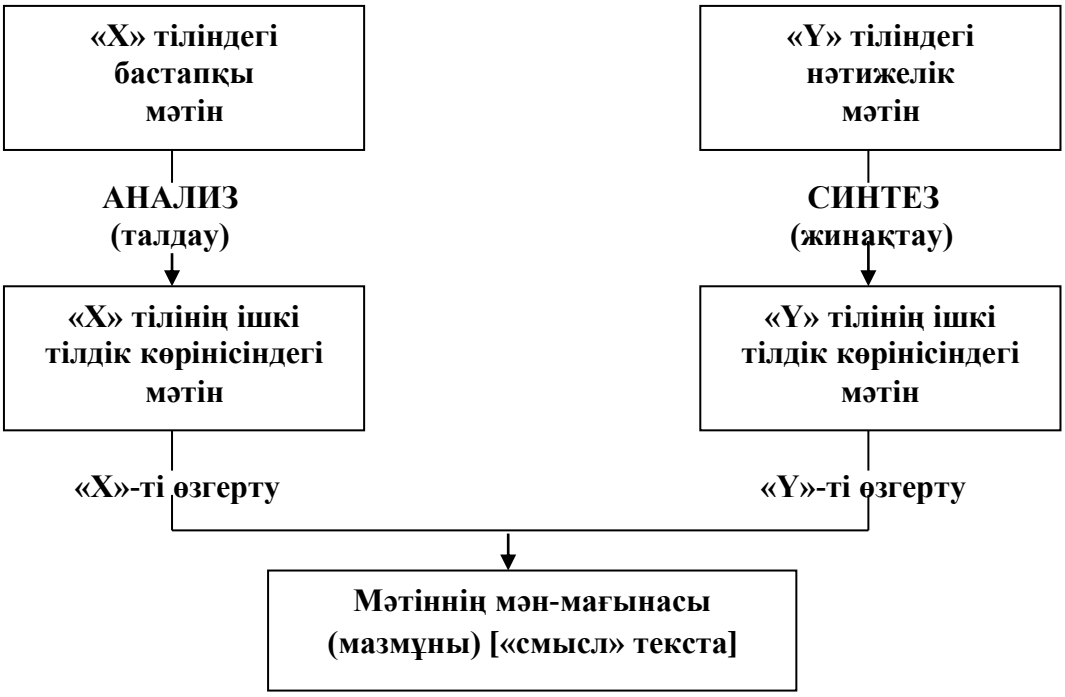
Машиналық аударма жүйелері кибернетикалық модель жасау аясына қатысты болуынан тілдік әрекеттің үшінші типі арқылы іске асатыны белгілі. Себебі аударма ісі электронды-есептеу машиналардың (ЭЕМ) көмегімен аудармашының әрекетін қайталауға бағытталады.

Аудармашының тілдік әрекетінің (сөйлеу-ойлау қызметі) моделін МА жүйесі арқылы іске асыру мынадай екі түрлі таптастыру параметрі бойынша анықталады:

- автоматтандыру дәрежесі, яғни компьютерге жүктелетін жұмыстың нақты көлемі. Бұл жерде мынадай жағдаяттарды бөліктеп қарастыруға болады. [Кулагина, 1979, 3-б.]: 1) аударма жұмысы толығымен (немесе толыққа жуық) ЭЕМ арқылы орындалады; 2) аударма жұмысын аудармашы атқарады да, ал ЭЕМ жәрдемші рөлін атқарады, мысалы, қажетті сөздің аудармасы сөздіктен тауып алу жұмысы;

- адамды (аудармашыны) қатыстыру стратегиясы, яғни аударма жұмысының негізгі бөлігі компьютер арқылы орындалады да, ал адам (аудармашы) аударылған мәтінді өңдеуге, яғни аудару кезінің әр тұстарында

(басында, ортасында, аяғында) редакциялық жұмыспен ғана айналысады [Machine Translation, 1987, 12-13 бб.].



1-сызба. АА процесінің идеалдық құрылымы.

Аудармашының сөйлеу әрекетінің (тәртібінің) ерекшелігі мынаған саяды: бастапқы нысан ретінде бір тілдегі мәтін алынады да ал тілдік қызметтің нәтижесі ретінде – мағыналық жағынан бірінші мәтінге барабар, басқа тілдегі мәтін болып табылады. Басқаша айтқанда, аталған нәтиже міндетті түрде мәтін күйінде ұсынылуы қажет. Ал басқа бір тілдік әрекет типтерінде нәтижені олайша дайындау міндеттілік сипаттама емес. Мысалы, бұлшық ет немесе ойлау әрекеттері болашақ қызметтің қажеттігіне қатысты білімді қорландыру және т.б. әрекеттер болуы міндетті емес.

Аталған ерекшелік МА-ға белгілі принциптік құрылымды сақтауды міндеттейді. Ол құрылым, автоматты аударманы (АА) жасаушының ұсынған нақты сызбасына тәуелсіз, міндетті түрдегі бөліктерге (компоненттерге) ие болуы қажет. Төменде ондай құрылым мүмкіндігінің ең жақсы деген (идеальный) құрылымы беріледі.

«X»-ке талдау (анализ) жүргізу мен оны өзгертудің мақсаты «X» тілінің сөздегі, грамматикасы және қолдану ережелері негізінде «X» тіліндегі мазмұнын ашу. Ал «Y»-ке синтез жүргізу – «Y» тіліндігі мән-мағынаның «Y»-ті өзгертудің, сөздіктің, «Y» тілінің грамматикасы мен қолданысың негізінде «Y» тіліндегі шығарылу мәтінін құрастыру немесе оны синтездеу.

ЭЕМ-сіз «адамша» аударма жасау кезіндегі талдау (анализ) мен жинақтау (синтез)компоненттерінің міндеттілігі айқын білінбегенмен, автоаматты аударма жасауда ондай бөліктер міндетті түрде орын алуы керек.

Ал «өзгерту» топтамасының қажеттілігі, тіпті, ең қарапайым жағдайда айқын байқалады. Мысалы, аударуға қажетті «X» тілінің сөздік бірлігі

мағыналық жағынан екінші тілдің («Y») сөздік бірлігімен сәйкес келіп, және екі тілдің де («X» пен «Y») синтаксистік құрылымы бір-біріне, тіпті сөздердің орналасу тәртібіне дейін тұспа-тұс келіп тұрған кезде де ешбір өзгеріссіз болмайтыны белгілі. Мұндай жағдайда өзгертуді қажет ететін лексикалардың грамматикалық формалары мен аударма баламасын өз орнына дұрыс қлю мәселесі.

Бірақ қарастырылған жайт идеалды жағдаятқа сай және тілдің шектеулі сөздік бөлігі мен синтаксисіне тән деуге болады. Сонымен бірге мұндай жағдаяттың тек әрбір нақты жұп тілдер үшін ғана ұтымды болуы мүмкін. Әзірше екі тілдің лексикалық бірліктері (ЛБ) мен синтаксистік құрылымдары бойынша басым көпшілік тілдерге жоғарыда аталғандай тұспа-тұс сәйкестік жайлы айту мүмкін болмай келеді. Оның бір себебі әр түрлі тілдердегі көп мағыналы сөздердің мағыналар көлемінің сәйкес келуі. Лексикалық сәйкестіктің бұзылуының ең ақырғы жағдайы қайсыбір тілдің баламасыз лексикасының болуы. Олар, әдетте, тұрмыстық бұйымдарды, міндетті сипаттайтын сөздердің екінші (аударуға қажетті) тілдің лексикалық қорында болмауы. Мұндай жағдайда лексикалық элементті сипаттау немесе калькілеу оның мәнін түсінуді қажет етеді, яғни мағынаға сәйкес келтіру үшін тиісті өзгерту жүргізуді қажет етеді.

Әр тілдердегі сөздіктердің сәйкес келу сипатының әр халықтың дүние танудағы ақиқат ортаны бөліктеуді тілдік деректермен бейнелеуге тәуелділігін Сэпир – Уорфтардың лингвистикалық салыстырмалық теориясының анықтайтыны белгілі.

Егерде аударма мәселесін, соның ішінде автоматты аударманы қоса санағанда, тіларалық сәйкестіктің бұзылуы тұрғысынан қарастырсақ, онда аударма мәселесінің мақсатын екі тілдің мәтіндерін мағыналық сәйкестікке келтіру деп ұғыну қажет болады. Яғни екі тілдің әр түрлі элементтерінің (сөздік, құрылым, сөздер тәртібі және т.б.) мазмұнды ашу құралдары ретінде сәйкес келмеу жағдаятын мағыналық сәйкестікке келтіру деп түсіну.

Мұндай сәйкестікке келтіру мәселесін шешу жоғарыда көрініс тапқан 1-сызбадағы топтамалардың кез келген МА жүйесінде орын алуын талап етеді. Сәйкестікке келтіру мүмкіндігі және бір тілден екінші тілге аударуға болатындықтан принциптік жақтарыбарлық тілдерге ортақ болып келетін мына қасиеттерге негізделеді: тілдің сызықтық және таңбалық сипаты (Ф. Де Соссюр), ал таңбалық қасиеттің негізінде (мағыналық тұрғыда) барлық халықтардың ортақтығы қызметінің таңбалау ортақтығы және бізді қоршаған әлемнің физикалық және элементтік ортақтығы жатқанын айта кету қажет.

Ма-ның өмірге келуіне, біріншіден XX ғ. Екінші жартысынан бастап әр елдердің (континенттердің) бірнеше тілдеріндегі ақпарат ағымының қарқындап даму (өсуі), екіншіден оларды ғылыми-техникалық прогресс үшін меңгеру қажеттігі, білімді аудармашылардың жетіспеушілігі және ондай аударма жұмыстарының көп қаражатты қажет етуі себеп болды. Әрі қарайға баяндалуымызды МА жүйелеріне шолу жасау (тіпті қысқашада) деп түсінгеннен гөрі, машиналық (автоматты) аударма жұмысының дамуы жайлы қайсыбір түсініктеме (көрініс) деуге болады.

Аталған әлеуметтік себептер аударма мәселесінің шешімін табуда оның ұтымды жолдарын іздестіруді қажет етті. Бұған ең алдымен аудармашының тілдік қызметінің ерекшеліктері ішінен шеттілдік мәтіннің құрылымын зерттеудегі таза формалды талдау жолын жатқызуға болады [Панов, 1958]. Мысалы, аудармашыға мәтіннің грамматикасы және лексиканың көмекші сөздері мен есімдік сөздері және т.б. осы сияқты функционалды грамматиканың элементтері белгілі болғанымен, ол мәтіннің салалық арнаулы лексикасы таныс болмаған жағдаятты қарастыруға болады.

Шеттілдік мәтінді аударудағы оның тұрпатынан (формасынан) мән-мағынасын ашу ерекшеліктері модельдеу нысаны ретінде, ең алдымен, аудармашының қызметін – іс-әрекетін алуға бірден-бір себепші болды. Сонымен шеттілдік мәтінді аударуда деректерін «шеттетіп», тілдің семантикасын арнайы талдамай-ақ, формалды тұрғыда аударма баламасын табуға мүмкіндік туындайды.

Автоматты аударма жасаудың бастамасы ретінде ресми түрде 1949 жыл саналды. Тілді формалды тұрғыда қарастырудан кейінгі екінші мүмкіндік деп саналатын – электронды-есептеу техникасының өмірге келуі. Бірақ МА-ны механикалық машиналар жолымен іске асыру идеясы аталған мезгілден көп бұрын бастама алатыны белгілі. Мысалы, А. Кахердің ойлап тапқан аударма «Yава таа» газетіндегі 1924 ж. 24-ақпанында шағын хабарлама ретінде жарық көрген болатын [Переводная машина в Таллине, 1962]. 1933 жылы Кеңес Одағының П.П.Смирнов-Троянскиге бір тілден екінші тілге аудару машинасын ойлап шығарған үшін арнайы патен тапсырылды [Переводная машина П.П. Троянского, 1959].

Аударманың лингвистикалық принципі бойынша екі немесе одан да көп тілдердің тілдік бірліктерін мағына жағынан бір-біріне сәйкес қоюды сәйкестікті анықтау механикалық тәсілдер арқылы іске асты. Электеронды-есептеу машиналарының жоқтығы мен елдегі патенттік жұмыс жүйесінің тұйықтық сипатынан аударма жұмысының жетістіктері әлемдік ғылымға тек XX ғ. 50-жылдарында ғана белгілі бола бастады. Аударма жасау бастамасымен тұспа-тұс келді деуге болады. Француз инженері Георгий Арцрунидің аударма үшін ойлап тапқан және 1933 жылы патенттелген жаңалығының тағдыры да аталғандай жағдайға душар болды. [Сочве, 1960].

Ең бірінші МА жүйелері АҚШ-та пайда болды. Тарихи тұрғыда қарастырғанда, МА жүйесін құрастыруға жалпы тілдік көзқарас криптография (шартты белгілермен жазылған құпия жазу) саласының теориясы мен практикасынан қабылданды деуге болады.

Екінші жержүзілік соғыс кезінде әр елдерде, әсіресе англия мен АҚШ-та хабарлардағы шифрды жай жазуға айналдыру (дешифровка) жұмысын тек «қолмен» атқаруда ғана емес, сонымен бірге машина арқылы әрекет етуде де көптеген нәтижелерге жеткен болатын. Мысалы, гитлерлік басқару орындарының шифрланған бұйрықтары мен хабарларын жай жазуға айналдыратын «Ультра» атты дешифровалды машинаның құрастырылғанын айта кетуге болады.

Лингвистикалық көзқараспен қарастырғанда кодталған хабарды жай жазуға келтіру дегеніміз, ол лексикалық қоры белгісі және анықталмаған грамматикалық ережелерге ие шет тілінде жазылған мәтінді аудару процесімен (жағдаятымен) барабар. Мұндай аударма жұмысы бір кодтағы жүйеден екінші кодтағы жүйеге ауыстыру процедурасымен (әрекетімен) бірдей деуге болады.

Маман-дешифровщиктердің пайымдауынша табиғи тілде екінші тілге аудару жұмысы жеңіл. Себебі, қазіргі тілдердің сипаттайтын көптегеен сөздіктері мен грамматикаларының негізінде тілдің өзіне тән ерекшелігі, лексикасы, морфологиясы мен синтаксисі алдын ала белгілі. Ал есептеу техникасы ондай сөздіктер мен грамматикаларды қалаған тәртіпте өз жадында орналастыра алады.

1946-1947 жж. ЭЕМ арқылы МА жасау идеясы Принстонның ғылыми-зерттеу институтында арнайы проблема ретінде талқыланады. (У. Уивер, А.Л. Бут және т.б.). Бұл идея бірізді шешімге ең алғаш дешифровка ісінің американдық маманы Уоррен Уивердің меморандумында көрініс тапты. 1949 жылдың 15-шілдесінде ғалым өз меморандумын лингвистика, есептеу техникасы мен дешифровка салаларындағы екі жүз мамандарға таратқан болатын.

У.Уивер МА-ның принциптік мүмкіндігін мүлде бөлек тілдердің ұқсас қасиеттерімен және әр тілдің өзіне ғана тән формалды сипатта болуымен байланыстырады. Жалпылама айтқанда, У.Уивердің пікірі бойынша, қытайша жазылған кітап – басқа емес, ол қытайша шифрланған ағылшынша кітап [Машинный перевод, 1957, 33-45бб.]

Сонымен МА –ның бөлек ғылыми бағыт ретіндегі даму кезеңі Уивердің мемарандумынан бастама алады. Осы ғылыми бағыттың өткен шағына талдау жасай келе, оны шартты түрде негізгі үш кезеңге бөлуге болады. [Hutchins, 1986; Кулагина, 1979; Марчук, 1983, 91-б.; Апресян және т.б., 1989, 4-5-бб.]. Мысалы, хронологиялық жағынан былайша бөліктенеді: 1-кезең: 1949 – 60-жылдардың ортасы; 2 – кезең: 60 – жылдардың ортасынан – 70-жылдардың аяғы; 3-кезең: 80-жылдардан қазіргі кезеңді қоса есептегенде. Мұндай үш кезеңге бөлуге негіз болған ғылыми-техникалық нәтижелер екені белгілі. Сондықтан лингвистикалық ,математикалық машиналық жабдықтаудың және практикалық нәтижелердің даму сипаты тұрғысынан қарастырғанда МА-ның бірінші, екінші және (қазіргі) «ұрпақтары» туралы жиі айтылады.

Уивердің меморандумымен жарияланған бірінші кезең 1949 жыл Вашингтон университетінде құрылған орталықтан (ағылшынша-орысша МА жүйесі) басталады, Сосын 1950 жылы А.Д.Бутаның жетекшілік етуімен французша-ағылшынша аударма орталығы құрылды. Совет Одағында бірінші аударма (машиналық) орталығы. Мәскеуде СССР ҒА-ның В.А.Стеклов атындағы Математикалық институтында О.С.Кулагинаның жетекшілігімен 1955 жылы ашылды.

Осы кезеңде ең бірінші үміттендіретіндей нәтижелер алынып, алғашқы машиналық тәжірибелер жүргізілген болатын. Мысалы, 1954 жылы АҚШ-та

Джорджтаун университетінде проф. Леон Е.Достердің жетекшілігімен GAT жүйесі негізінде құрастырылған орыс тілінен ағылшын тіліне аударатын, МА жұртшылық назарына ұсынылды (демонстрацияланды). Бұл бір тілден екінші тілге ЭЕМ (IBM-701) арқылы орындалған ең алғашқы аударма тәжірибесі болатын. Ғылымда бұл тәжірибе Джорджтауындық эксперимент атпен тарихта қалды [Машинный перевод, 1957, 171-б.]. Аталған экспериментте «тікелей» МА жағдаяты (бір тілдік жұпқа қатысты жүйелер) тексерістен өткен болатын. Жүйе сөздігі барлығы 250 сөзді қамтыса, ал оның грамматикасы болғаны алты синтаксистік ережеге негізделеді [Machine Translation, 1987, p. 22-23]. Совет одағында француз тілінен орысша аударатын МА-ның бірінші эксперименттік жүйесі О.С. Кулагина мен И.А. Мельчуковтың авторлығымен 1955-1956 жж. Құрылды [Кулагина, Мельчук, 1956]. Бұл жүйедегі негіз сөздер сөздігі 1236 сөз, ал сөз айналымы сөздігі 250 бірлік қамтиды. Аударма бірлігі ретінде толық сөйлем алынды. Морфологиялық талдау (анализ) алгоритмі жалғаулардың кестесімен әрекет ету арқылы аударылатын фразаның (сөйлемшенің) сөздеріне қажетті деген грамматикалық ақпаратты тіркеп отырады. Бұл жүйенің артықшылығы омонимдік сөздерді ажырата алатын топтамасы (блогі) болуында орыс тіліндегі фразаны синтездеу топтамасында сөздердің орналасу тәртібінің дұрыстығын тексеру мүмкіндігі алдын ала қарастырылды. Сонымен бірге Т.Н. Молошнаяның жетекшілігімен ағылшынша-орысша аударма жасау жүйесінің алғашқы жобасы ұсынылды [Молошная, 1960].

Жаңа бағыттың қарқынды дамуы көптеген мәртебелі ғылыми конференцияларда, симпозиумдарда, жарияланымдарда көрініс тапты және МА проблемасымен айналысатын бірнеше ірі орталықтар ,топтар мен зертханалар ашылуына себепші болды. Өткен ғасырдың 60-жылдардың басында олардың әр елдердегі сандық көрсеткіші мынандай: ФРГ - 22, АҚШ - 21, СССР - 10, КНР мен Жапония – 6-дан, Франция – 5, Чехославакия – 4, Великобритания мен Венгрия – 3-тен, ГДР мен Италия – 2-ден. Осындай топтар Бельгияда, Болграияда, Израилде, Канадада, Люксембургте және т.б. елдерде пайда бола бастады [Автоматический перевод, 1967, 509-б.].

МА жұмысының бастамасы жалпы алғанда өрлеу сипатында болды – тілшілер зерттеудің жаңа бағытын ашты, ал програмистер күн тәртібіне күрделі және аса қызықты мәселелердің шешімін табуға бел буды. Бұл кезеңде тіл мен аударма процесіне қатысты барлық мәліметтерге қол жеткен сияқты көрініп, тек олар формалжы пішінге келтірілсе, МА-ның әмбебаптық жүйесі құрылады да негізгі мақсатқа тез арада жететіндей болып көрінген. Бірақ алғашқы және одан кейінгі тәжірибелердің нәтижелері негізінде байқалған қиындықтар бұл мәселенің шешімін табу ойлағандай оңай еместігін байқатты.

Солай бола тұра бұл жұмыстар тек МА аясындағы зерттеулерге ғана деп беріп қоймай, тілдің теориялық және құрылымдық жағына да көңіл аударуға ынталандырды.

Қорыта айтқанда, автоматты аударманың кезеңі ерекше (айтарлықтай) жемісті болды. Бұл кезеңде көптеген іргелі мәселелер ірге тасы қаланып,

зерттеудің болашағы сараланып, жолдары анықталады. Мысалы, автоматты сөздіктердің құрастыру, аралық тілді (АТ) жасау (создание), сөздердің омонимдік сипаты мен синтаксистік құрылым, яғни сөздердің сөйлемдегі тіркесімі жайлы, фразеологизмдердің жасалу жолдары және көптеген басқа мәселелердің сыр-сипаты анықталды.

Осы мезгілдерде, бүгінгі таңда да аса маңыздылығын жоғалтпаған, синтаксистік талдаудың алгоритмінің бірінші нұсқалары зерттелді: синтаксистік тәуелділік пен тікелей құрастырушыларды, болжамдық талдауды, көп вариантты синтаксистік талдауды, «фулькрумдер» әдісімен талдау, конфигурациялық талдау және т.б. ескере отырып талдау жүргізу.

Терминтану саласындағы тәуелділіктің моделі мен тікелей құрастырушылардың моделін талдау бүтін сөйлемнің құрылымдық байласымын анықтауға болатындығына ұйғару жасайды. Мұндай құрылымның көп-мағыналығын басын ашып алу үшін әр түрлі «сүзбелердің» әдістері қолданылады (мысалы, мына жұм. қара: ЛМУ-дың математикалық лингвистика зертханасындағы С.Я. Фитиалов пен Г.С. Цейтиннің жетекшіліктерімен жүргізілген жұмыстар; GETA атты Гренобльдық топ – Centre d’études pour la tradution autmatigue, жетекшісі В.Вокуа). «Күшті» сүзбе ретінде Д.Е. Хейстің байланыстың «күші» ұғымы ұсынылды (The Rand Corporation Automatic Language Data Processing Coroup. Santa Monica (Calif.)). Талдау варианттарының артықшылығы бар деп қайсысының байланыс күші жоғары болса сол түрі саналады [Автоматический перевод, 1971. с. 41].

Болжамдық талдау (Национальное бюро стандартов США, И.Родес; Вычислительная лабаратория Гарвардского университета (Масс.), рук. А.Эттингер) сөздік мақаланың айырықша құрылымына негізделеді. Мұнда әрбір сөзге оның барлық мүмкіндік байланысы туралы ақпарат тіркеледі. Талдау тәртібі (процедурасы) мәтінді солдан оңға қарай тексеріп шығуға негізделеді. Осының нәтижесінде әрбір кезекті (келесі) сөз қайсібір өлшемі бойынша алдыңғы сөз негізінде жасалған болжамға сай келуі тексеріледі. Егер жасалған болжамға сай келмесе, ондай сәйкестікті іздестіру солға қарай ауысады. И.Родестің ғылыми тобында бірварианттық талдау қарастырылды, яғни талданып жатқан сөздің сол жағында орналасқан бірінші сөз – сәйкестік «иесі» деп саналып, кез келген сөйлемге тек бір ғана синтаксистік құрылым сәйкестікке қойылған.

Сөйлемнің синтаксистік құрылымының бірнеше варианттарын алуға мүмкіндік туғызатын көпварианттық талдау бір қадым алға басқан болып табылды (С.Куно, А.Эттингер, Гарвард университетінің есетеу зертханасы). Бұл көпварианттылық сөздің синтаксистік потенциалды мүмкіндіктерінің көптігінің арқасында синтаксистік байланыста тұратын сөздерге болжам жасай алатындығына негізделеді. Көпварианттылықтың талдаудың алғашқы грамматикасы 2100 ереже мен 82 синтаксистік болжамды қамтыды. Сөйлемнің құрылымының диаграммасы тікелей құрастырушылар моделі негізінде құрылды. Осының нәтижесінде мүмкін болатын барлық варианттар іріктеліп, олардың көпшілігі әр түрлі сүзбелер көмегімен сан жағынан азайтылыды. Болжамдықталдау алгоритмінде көпварианттылығы фразаның

синтаксистік құрылымына тән екі мәнділік ақиқаттығынан гөрі, ол формалды және лингвистикалық шектеулерді жеткілікті түрде зерттеп алмай тұрып, таңдалып алынған варианттарды іріктеу әрекетінің формалды сипатының салдары деуге болады. Міне бұл жүйелердің осал жерлері осында. Дегенмен синтаксистік болжам жайлы идеяның (ой, пікір) өзі құнды және оған грамматикалық болжамдарға сүйенетін кейбір қазіргі кездегі автоматты талдау жүйелері негізделеді. [Цейтин және т.б., 1982; Откупщикова, 1982].

Атап кететін жайт, ол бірварианттық пен көпварианттық талдауларды бір-біріне қарсы қою МА-ның қалыптасуының бірінші кезеңі үшін ерекше актуалды. Бірварианттық талдауда кездейсоққа жақын дәрежеде таңдалып алынған сөйлем құрылымы дұрыс болмай шығуы мүмкін, ал талдауды басынан қайта бастау мүмкіндігі мұндай жүйелерде қарастырылмаған. Көпварианттық талдауда біркелкіліктің бұзылу қаупі бар: құрастырылған құрылымдардың саны дұрыс вариантты таңдау мүмкіндігінен айтарлықтай үлкен болуы мүмкін.

Сонымен, бірге жақсы аударма шы (немесе аударма жүйесі) аударуға тиісті фраза енетін барлық мәтінмазмұнынтүсіну нәтижесінде, оның дұрыс түрдегі бір ғана вариант ұсынады. Ал фразалардың көпмағыланығықисынды (байласымды) мәтін бойынша өз шешімін табады. Бірақ, бұл жағдайда, семантикалық мәліметтер, мәтінде баяндалатын ақиқат жағдаяты (тіл біліміне қамтамасыз білім) және, ең соңында, толық мәтіннің синтаксистік пен семантикалық заңдлықтары кең қолданыс табады. Мұндай мәліметтер МА-ның бірінші және екінші кезеңдерінде толығымен қолданылмады, фл қолданылса да фрагменттік (үзінді) түрде ғана көрініс тапты деуге болады. Мұндай жағдайға душар еткен қажетті бағыттағы лингвистикалық жабдықтау мәселесінің толық зерттелмеуі негізгі себеп болып отыр.

Конфигурациялық талдау аударуға тиісті тілдің мәтінін (аударма жүйесіне енетін мәтін) сол тілдің алдын ала бөліп алынған синтаксистік конструкциялар (конструкциялардың, формулалардың) жиынымен салыстыруға негізделеді [кафедра математической лингвистики ЛГУ; Белопольская, Ковригин, 1963; Воронин, 1961. Кафедра структурной и прикладно лингвистики МГУ: Дрейзин, Рашитов, 1961. Группа автоматического перевода ИЯАНСССР: Молошная, 1960 и др.]. Егер аударма жүйесіне енетін мәтіәннің анықталған конфигурациясы жиынның синтаксистік конструкцияларының (конфигурациялардың) бірігуімен сәйкес келіп жатса, мәтіндік конфигурация танылды деп саналады да, ол «ықшамдалады» («свертывается») және әрі қарайғы талдау әрекетінде ықшамдалған немесе «қысылған» түрінде көрініс табады. Синтездеу (жинақтау) әрекетінде талдау кезінде анықталған конфигурация оған сәйкес келетін аударуға тиісті тілдің конфигурациясымен салыстырылады. Жүйеде конфигурациялардың сәйкестігі алдын ала ескеріледі (қойылдаы - устоновлено).

«Фулькрумдер» әдісімен талдау-мынаған негізделеді: сөйлем төменгі сатыдағы компоненттерге (блоктарға) бөліктеледі. Олар фразаны жүйелі түрде бірнеше рет қарастырудың (циклдық қайталаулар) негізінде одан ірі

фрагментерге біріктіріледі. Әрбір кезекті қарастыруда талдаудыңтірек нүктелері (fulkra) анықталады. Міне осындай фулькрумдар синтаксистік конструкциялардың негізгі және «басқарушы» сөздер қатарына жатады. Бұлардың құрылымы мен шекарасын анықтау циклдық қайталаулардың мақсаты болып табылады [группа автоматического перевода фирмы Bunker Ramo-Woldridqe Inc, Canoqa Park, Calif., рук. П.Гарвин].

Тізбекті (цепочечный) талдау әдісі арқылы фразаның синтаксистік құрылымын тану үшін берілген сөйлем көрінісіндегі синтаксистік категориялар тізбегін элементарлық тізбектерге бөлу қажет болады. [З.Харрис, Н.Сейгер].

Корреляциалық талдау кезінде әрбір сөзге индекстер жиынтығы тіркеліп жазылады. Олардың әрбіреуі сол сөздің осыған ұқсас индексі бар басқа сөздермен өзара байланыстылық (корреляция) мүмкіндігін бейнелендіреді. Корреляцияның (өзара қатынастылық) жоғары деңгейі лингвистикалық шектеулер арқылы төмендетіледі. Мысалы, сөз, семантика және т.б. сатыларға төмендейді.

Осымен бірге зерттеушілер МА мәселесі үшін толық (бүтін) мәтінді талдау проблемасын қарастыра бастады. Сонымң ішінен бүтін (толық) мәтіндердің антецедент-зат есімдермен арақатынасынанықтайтын ережелердің қарастырыла бастағанын атап кетуге болады. (антецедент-анафоралық қатынас. Мәселен, есімдіктен бұрын тұратын есім сөз. Ал есімдік осы есімге сілтейді) [Шумилина, 1961; Автоматический перевод, 1967, с. 235 к сл.].

Қазіргі кездегі жобаланып жатқан МА екі класқа бөлінеді: бір тілден екінші тілге аударма жасайтын бинарлы және бірнеше тілдерден бір тілге немесе бірнеше тілдерге аударуға арналған – көптік (множессвенные).

МА жүйелерінің бірінші түрінің әр уақытта басымшылық сипаты адам баласының аудару мүмкіндігін модельдеудің және әрбір жұп тілдер деректерінің бинарлық сәйкестіктерін анықтаудың теориялық және практикалық мүмкіндіктерінің табиғилық жағдаятынан туындайды. МА-ның бинарлық жүйлерінің кемшілігі ретінде аударуға тиісті керістік (входной) тілді талдауда олардың қасиеттеріне қатаң тәуелділікте болу сипатын айтуға болады. Бұл тәуелділік аударма алгоритмінің қатаңдығына әрі қарай қиындай түсетіндігі байқалады.

Көптік аударма жүйесі аталған қатаң тәуелділіктен құтылуға (сақтандыруға) мүмкіндік тудырды және талдау мен жинақтаудың (анализ бен синтездің) тәуелсіз алгоритмдерін құрастыруға жол ашты. Бинарлық аудармалардың қосындысы тұрғысынан түсіндірілетін көптік аударма олардың саны n(n-1) формуласы арқылы анықталады. Бұл өрнектігі «n» аударма процесіне қатынсытын көптік тілдер саны.

Көптік жүйлерді зерттеуші ғалымдар талдау мен жинақтаудың баламысы (инвариантты) ретінде аралық-тілді (интерлингваны) ұсынды [Машинный перевод, 1957].

Көптік аударма жасаудағы «аралық-тіл» идеяны іске асыруда бірінші еоеулі үлес ғалым Н.Д.Андреевке қатысты [Андреев, 1957]. Одан кейін

ЭЛМП (экспериментальная лаборатория машинного перевода – Ленинград, ЛГУ, рук. Н.Д.Андреев) қызметкерлердің осы бағыттағы көптеген ғылыми еңбектері жарық көрді. Аралық-тіл (АТ) мәселесіне қатысты Кеңес Одағы мен шет елдердің ғылыми орталықтарымен басқа да топтарында аралық-тіл (АТ) мәселесіне қатысты ой-пікірлер қызу талқылана бастады.

Н.Д.Андреевтің концепциясы бойынша, АТ ұғымының негізінде табиғи тілдерде конгруэнттік сипаттың болмайтындығы жайлы түсінік жатыр. Яғни АТ-ның жасандылығының заңды екендігі, сол себепті одан конгруэнтті емес элементтердің (артиклдер және т.б.) алынып тасталуы мен көпшілдік тілдер үшін конгруэнтті (етістіктің үш түрлі шағы, көптік жалғау және т.б. тәрізділер) сақталып қалуы қажеттігі туындайды. Сондықтан, аталған мәселеге арналған алғашқы зертеулерге аралық-тіл (АТ) өзінің құрылымы жағынан табиғи тілге жақын тұрып, соның терминдерімен баяндалады.

Әр тілдің бірліктері бойынша грамматикалық мағыналарының, грамматкиалық категориялары және бәріне ортақ семантикалық мағыналарының салыстырмалы сипатының қорытындылаушы қасиеттерінің белгілі болуына қарамай, семантикалық теорияның қолданбалы тұрғыдай жеткіліксіз зерттелуі тілшілерге үйреншікті тілдік көзқарастан ауытқуға кедергі жасады [Мельчук, 1959; Автоматический перевод, 1967].Н.Д.Андреев басқарған машиналық аударманың тәрірибелік зертханасында (ЭЛМП) ЭЕМ-ның көмегімен 18 тілден орыс тіліне аударатын және аралық-тіл (АТ) идеясын жария ету мақсатындағы шағын дәрежедегі тәжірибе жүргізілді.

Ат идеясын өмірге әкелу мақсатындағы сан алуан әрекеттерге қарамай, көптеген тілдерден бір тілге аудару тәжірибесі МА тарихында айтарлықтай өз орнын алмады деуге болады. Теориялық тұрғыда қарастырғанда оның себебі аралық-тілдің (АТ) грамматика мен лексика терминдерінде сипатталуында деп деп түсіндіріледі. Шындығында, ол тек қана семантикалық тұрғыда, семантикалық барабарлық түсінігінде көрініс табуы қажет болатын.

Сыншылар АТ арқылы аударма жасаудың төмен дәрежедегі барабарлығын айта келе, оның аударуға тиісті мәтіннің маңызды сипаттамаларының назардан тыс қалып қоятындығын айтады: мәтін авторының коммунативтік және праграматикалық ұстанымы, талдаулық (бөліктердің) актуалдылығы және т.б. Бұл сол кездегі ұсынылған АТ моделіне қатысты сөзсіз болатын жағдай деуге болады.

МА жүйесіндегі АТ-ның жүзеге аспай қалуын лингвистикалық зерттеулердің практикалық көзқарасымен түсіндіретін болсақ, ол өзгеріс енгізу этаптарының әдеттен тыс күлделілігінен, ал техникалық жағынан алғанда ЭЕМ-ның алғашқы буындарының көлем жады мен әрекет жылдамдықтары көрсеткіштерінің төменгі дәрежелігі деуге болады. Мұндай кедергілер бинарлы аударманың экспериментті жүйесін құру жағдайында да теріс әсерін тигізді деуге болады.

Қазіргі кезде аталған кедергілер әр түрлі АТ-ның іске қосу кезінде өз шешемін тауып отыр. Аралық-тіл мәселесіне қызығушылықтың екеінші айналымы жасанды интелектің жүйесін құрастыруға байланысты болды.

Себебі, тек осы идея бойынша ғана қазіргі кездегі кибернетика саласында зерттеліп жатқан мәтіннің мазмұнын түсінетін интелектуалды жүйелерге жол ашылды. Бірақ, бүгінгі таңның өзінде-ақ аралық-тіл арқылы машиналық аударма жұмысын іске асыру кез келген аударма жүйесін құрастырушының арманы, жоғарғы мұраты деуге болады. Солай бола тұра, аралық-тіл мәселесін зерттеу семантикалық теорияның, табиғи тілдің фразалары мен мәтіндердің семантикалық көрінісін анықтау жағдаяттарының даму процесіне зор ықпалын тигізді.

МА-ның дамуына аралық-тіл концепциясының күмәнсіз үлесі аударатын қасиеттіріне ғана негізделетін анализ (талдау) бен синтездің (жинақтау) бір-біріне тәуелсіз екендігін мойындау; аударылатын сөйлемнің мән-мағынасының үйлесімділігі мен әмбебептық берілісін зерттеу қажеттігін мойындау болып табылады.

МА жүйелерін құрастыру аясындағы жүргізіліп жатқан жұмыстардың белсенділігіне қарамай, бірінші кезеңде қалағанындай практикалық нәтижеге жете қойған жоқ. Аударманың тұтынушы сұранысын қанағаттандыра алатын өндірістік жүйесін құру жұмысының нәтиежесіз болуының үлкен зардабы байқалды. Мысалы, зерттеулердің бір бөлігі МА мәселесін шешуден бас тартты. Осы бағыттағы жұмыс істейтін бірнеше топтарының зерттеу әрекеттінің тоқтап қалуына себепші болған АҚШ-тың ұлттық ғылым академиясының арнайы комиссиясының әуе соғыс күштерінің МА-сы жайлы жасалған талдаудың қорытындысын бұрмалап түсінудің (экстраполяция) де өз рөлі болды. Осының зардабынан көпетегенғылыми ұжымдар ұйымдастыру мен финансалық қиындықтарға душар болды [Кулагина, 1979, 14-б.].

Мұндай жағдайға душар болуға кейбір ғылымдардың скепсистік көзқарасы, МА идеясының іске асу мүмкіндігіне сенімсіздік білдіруі де, әсіресе ғылыми-техникалық мәтіндерді аудару мүмкіндігіне күмән туды. Ал көркем әдебиет мәтіндері, оның ішінде, әсіресе поэзия мәтініне байланысты МА жасауға ең басынан-ақ ғалымдар қарсылық пікір білдірді. Оған шығарманың көркемдік, бейнелік жағын ұстаным етудегі кейбір тілдік заңдылыққа қайшы келу жақтары негізгі себеп болды.

Осындай жағдайды байланысты И.Бар – Хиллелдің көзқарасына тоқтала кетуге болады. Ғылым жоғары сапалы толы автоматтандырылған МА мүмкін емес және утопия деп санайды. И.Бар – Хиллелдің пікірінше МА кезіндегі негізгі кедергі сөздердің көпмағыналы және тілдің өзінде ондай екі мәнділікті ажырата алатын формалды көрсеткіштер жоқ. Оның шешімі тек әлем туралы білімде деп тұжырым жасайды [Бар-Хиллел, 1962].

Жоғарыда айтылған, сенімсіздік пікірлер әлем бойынша МА мәселесімен айналысатын топтар санын азайтуға соқтырып, екінші кезеңнің аяғында 100 топтан 38 топ қана қалғанын айтуға болады [Кулагина, 1979, 16-18 бб.].

Бірақ МА-ның болашағына сенімді, дәйекті көзқарастағы қажымай-шаршамай жоспарлы ізденістерін жалғастыра берді. МА-ның екінші кезеңі басталды. Енді мамандар зерттеудің теориялық жағын тереңдете түсіп, формалданған лингвистикалық модельдер жасауға, жүйелі және

функционалды грамматика жазу мәселесіне кірісті. Бұл кезеңде семантиканы моделдеуге арнайы назар аударылды. Бірінші кезеңнің нәтижелерісараланып, болашақтағы прагматикаға бағышталған жүйе құрудың мүмкіндігі анықталды.

Бірінші кезеңде зерттеле бастаған формалды грамматикаының модельдері: теоретика-көптік модельдер, тәуелділіктер грамматикаларының модельдері, тікелей құрастырушылар, туындату грамматикасы мен семантикалық модельдер және т.б. Екінші кезеңде машиналық аударма (МА) мәселесінің зерттелуі жалғасын табады, интерпретация (түсініктеме) беріледі, жеке тілдердің грамматикалық құрылымдарының азды-көпті толық түрдегі сипаттамаларында қолданыс табады. Сонымен бірге автоматты талдаулардың іргелі алгоритмдерінде апробацияланады (мақұлданады).

Екінші кезеңге келесі жататындар: семантикалық талдау принциптерін бұрынғыдан қарқынды зерттелуі және жоғары сапалы аударма жасауға қатысты талдаулар жасау жүйелерінде семантиканы қатыстырмайынсыз болмайтындығын мойындау. Соңғы айтылғанға байланысты – қайсыбір жүйелерде талдаудың айырықша семантикалық деңгейін (деңгейлерін) енгізу арқылы (мысалы, ФРАП), ал екіншілерінде – алдын ала бергені синтаксистік қатынастар типтерінің кең түрдегі синтаксистік сипаттағы тізбесі арқылы (мысалы, «мән-мағына – мәтін» моделінде). Үшіншісінде – аралық-тіл (АТ) типтегі мән-мағыналық инвариантын ұсыну арқылы және осыған келтіруге болатын семантикалық ережелерді келтіру арқылы және т.б. Сонымен бірге екінші кезең МА жүйесін құруға байланысты жұмыстардың жалғасуымен және оларды тәжірибелік сынау мен өндіріске енгізу әрекеттерімен сипатталады.

Сөздік пен базалық алгоритмдардің көлемдерін ұлғайту арқылы Джорджтаундық жүйенің (GAT) және олардың дериваттарының, жекелеп айтқанда SYSTRAN жүйесі дамыды. 1976 ж. Бұл жүйенің ағылшын-французбдық версиясы Европалық экономикалық бірлестік қоғамына тапсырылған болатын. Аталған жүйеде негізі синтаксистік байланысты тағайындайтын (устанавливающий) синтаксистік талдау ескерілген [Марчук, 1985]. Анализ бен синтездің айтарлықтай тәуелсіз болуынан оларды басқа қатарлас тілдік жүйелерге де қолдану мүмкіндігі туады (апплицированы).

1971 жылдан бері зерттеліп келе жатқан орысша-французша аударманың Гренобльдық жүйе (GETA, Б.Вокуаның жетекшілігімен) жетілдірілген түрінде логикалық пішіндегі терең синтаксистік құрылымдары алуды көздейді. Бұлар бойынша семантикалық, синтаксистік, морфологиялық және сөздік деректері арқылы бірнеше өзгерістер (трансформаций) енгізе отырып, шығу тетігіндегі тілдің (выходной язык) үстірт тұрпатының көрінісін (поверхностная форма) анықтауға болады. Көпдеңгейлік талдау өңдеудің 11 топтамасының (блогінің) реттік тәртібі түрінде құрылған. Мысалы, аналитикалық тұлғаны синтетикалыққа келтіру, фразаны (айтылымды) сегменттеу, есім топтарын т.б. Мұндай әрекеттер антецеденттерді іздестіру мен өалпына келтірілген түйісу нүктелердің (восстановленных узлов) синтаксистік және логикалық фунцияларды

анықтауға дейін жалғасады. Бұл жүйе ARIANE атты дамыған программдық (бағдарламалық) жабдықтауға (обеспечение) ие. Мысалы, ол жүйе мен тұтынушы арасындағы әрекетті деректер базасы арқылы алдын ала қарастырумен қатар процесті басқару мен мәтінді өңдеуде де дамыған құралдарды қолданады.

TAUM жүйесінде (монреаль университеті, Канада, жетекші – Р.Китредж) ағылшын-французша аударма үстірттік (поверхностный) синтез деңгейінде іске асады [синтаксический компонент ..., 1981]. Аталған жүйе лексикалық және құрылымдық жағынан өзгерту (преобразовать) міндетін атқаратын трансфер топтамасына (блок) ие: лексикаларды аудару французша фразаны синтездаудің жеңілдету үшін құрастырушылар тармағының (дерево) кейбір бөліктерін өзгерту, етістіктің шақтық түрін және ... моделін (модели управления) аудару мәселелері. Программа жазу тілі ретінде А.Кольмер құрастырған Q (Quebec) тілі алынған [синтаксический компонент..., 1981, с. 92]. Бұл жүйедегі талдағыштың комбинаторлығы мен икемділігі (талдаудың алғашқы болжамы аяғына дейін сақтау, ережелердің қолдану тәсілдерінің ретін белгілеу және т.б.) аударманы үстірттік синтаксискебағышталуына мүмкіндік тудырады.

TAUM жүйесі 1968 жылдан бастап 1980 жылға дейін белсенді дамыды, 1977 жылдан бастап жылына 5 млн. сөз көлеміндегі мәтіндерді аударып, ол өндірістік жүйе ретінде қызмет етті [Machine Translation6 1987, p. 12].

Чехославакияның Карлов университетіндегі алгебралық және есептеу лингвистика тобы П.Сгаллдың жетекшілігімен жай сөйлемдердің ағылшын тілінен чехия тіліне шектеулі сөздік бойынша аударма жасайтын МА жүйесі құрастырылады [Синтаксический компонент..., 1981, с. 84]. Талдау процесі негізі төрт кезеңге бөлініп қарастырылған: морфологияға дейінгі талдау, морфологиялық талдау, аудару. Синтаксистік талдау синтаксистік ережелердің әрекет ету аясының кеңейтілген типі ретінде құрылған: алдымен шағын мәнмәтіннің бірліктері талданады және әрі қарай көлемі біртіндеп ұлғая отырып, сөйлем шеңберінен шығатын мәнмәтінге дейін өседі. Сондықтан, мұны әдеттегі «төменнен жоғары» қарай талдау жағдайына жатқызуға болады. Осыған байланысты синтаксистік талдаудың өзі бірнеше этапқа бөлінеді. Аталған жүйеге жүйесіз (нерегулярный) синтаксисті, мағынаны дұрыс ашуға қажетті сөйлемде жоқ элементтерді толықтыру ережелерін, талдау үшін «авариялық» грамматика енгізілген. Синтаксистік қатынастарды семантикалық семантикалық түсініктеме беру тәсілі кең түрде қолданыс табады.

Шамамен алғанда 1970 жылдан SUSY жүйесін құрастыру жұмысы басталады [Синтаксический компонент..., 1981, с. 90]. Аударма жасау процесі деректерді алдын ала дайындаудан, талдаудан, өзгеріс енгізуден (аудару) және синтездеу (жинақтау) тұрады. Аударма алгоритмін жасау әр тілдердің грамматикасына тәуелсіз принципіне негізделеді, сондықтан жүйеге жаңа тілді қосу әр түрлі жаңа бағдарламаларды құрастыруды сан жағына шектейді. Неміс, орыс, ағылшын және француз тілдерін талдау, ал сонымен бірге неміс тіліне қатысты синтездеу (жинақтау) жұмыстары жүргізілуде. Талдау

жұмысының нәтижесі ретінде туындау тармағы (дерево) түріндегі «базистік құрылым» болып саналады. Сөздіктер жүйесі (талдау сөздіктері, семантикалық сөздіктер және аударма сөздіктері) әсіресе талдау процесіндегі семантикалық ақпарат үшін аса маңызды болып келітн кез келген ақпаратты пайдалануға мүмкіндік туғызады. Талдау принципі – құрастырылған (қиыстырылған): талдаудың бірінші бөлігі «жоғарыдан төменге қарай» өңдеуден басталады, екінші бөлігі – «төменнен жоғарыға қарай» өңдеуден басталады, екінші бөлігі – «төменнен жоғарыға қарай», үшіншібөлікте алдыңғы екі нәтиже бір-біріне «жапсырылады». Бұрыс синтаксистік құрылымды өңдеу үшін талдау соңында семантикалық талдау процедурасын іске қосу алдын ала қарастырылған.

Ресейлік МА-мен айналысатын топтардың сан жағнан азаюына қарамай, біз қарастыратын екінші кезең зерттеу жұмыстарының нәтижелігімен сипатталады. Оларды былайша атап кетуге болады:

а) лингвистикалық және математикалық жабдықтау тұрғысынан алғанда, бұрынырақ зерттелген жүйелердің жақсартылған түрлерінің құрастырылуы;

ә) жаңа типті жүйелердің құрастырылуы (Бүкілодақтық аударма орталығы, А.И.Герцен атындағы ЛГПИ-дегі «Статистика речи» тобы);

б) автоматты талдау жүйелерін қарқынды түрде зерттеледі;

в) МА жүйелерін әр түрлі мәндік дәрежеде тәжірибелік сынауға дейін жеткізу.

Соңғы бапта осы кезеңге дейінгі құрастырылған жүйелердің жетістігі мен кемшіліктері бірдей дәрежеде сыннан өтеді.

Осы кезеңде ССРО ҒА-ның қолданбалы математика институтында француз тілінен орысшаға аударатын автоамтты жүйенің жетілдірілген жүйенің жаңа түрі (версиясы) құрастырылды. Француз тіліндегі математикалық мәтінді орысшаға аударатын бұл жүйе 1967-1976 жж. Бойы О.С.Кулагинаның жетекшілігімен зерттеу жұмыстарымен шұғылданады [Кулагина, 1979]. Жүйеде аударма әрекетіне қатысты өзгеріс былай іске асады: жүйеге енетін ақпаратфранцузша фразаның үстіңгі синтаксистік құрылымы (фраза бойынша аудару), ал жүйеден шығарды – мән-мағыналық сәйкестіктетұратын орысша фразаның үстіңгі (поверхностный) синтаксистік құрылымы. Мұндай тәсілдің негізінде екі тілдің сөздер мағыналарының сәйкестігі және фразалардың мағыналары мен құрылымдық сәйкестіктері жатыр. Бұлайша талдаудың мақсаты тәуелділіктер таралымы (дерево) түріндегі фразалардың синтаксистік құрылымдарының жиынын құру болып табылады. Осылайша синтаксистік байланыстың 52 типі бөлініп алынды.

Көпварианттық талдауды сүзбе-ережелер жиыны қолданылады. Ол арқылы синтаксистік байланыстың күші мен қашықтығы байланыстың ортақтастығы немесе ортақтас еместігі (совместность / несовместность связей), құрылымдардың жобалылығы немесе жобалауға болатындығы (проективность / непроективность структур). Қолданылу ретіне қарай, сүзбелердің иерархиясы енгізілген. Мысалы, алымен қарапайым сүзбе, сосын жалпы, күрделі, соңында жекеленген жағдайға сай келетін сүзбе

қолданылады. Мұндай сүзбелер арқылы болажамдар жиыны ішінен көпвариантты талдау түрлеріне сүйене отырып, бір болжамы құрастыруға мүмкіндік туындайды [Кулагина, 1979; Синтаксический компонент..., 1981].

Орыс тіліне синтаксистік талдау жасауды қарастырғанда (зерттеуде) МА-ның даму кезеңінің алғашқы кезеңінде ұсынылған конфигурациялық талдау идеясы қолданыс тапты. Осының нәтижесінде сөйлем ішіндегі сөздердің синтаксистік қатынастырның әр алуан типтерін мағыналық жағынан сипаттайтын тікелей доминация қатынасының (ТДҚ) 31-і бөлініп алынды (ОНД- отношение непосредственный доминация). Синтаксистік талдаудың мақсаты аударуға тиісті сөйлемнің сөздері арасындағы синтаксистік байланысты анықтау. Ол үшін мұндай байланыстарға қажетті ТДҚ-ны тіркеп жазылады және берілген мәтінге оның мазмұнын бейнелейтін ерекше сипаттамалар сәйкестендіріледі. Автоматты талдау жүйесінің автоамтты аудару, реферат алу, ақпаратты іздестіру сияқты қолданбалы мәнімен бірге, ол теориялық жағынан да аса маңызды екенін атауға болады. Себебі «алгоритм анализа является действующей и экспериментально контролируемой моделью «анализирующей части языка»» [Автоматический перевод, 1967, с. 226; Мельчук, 1974].

ЛМУ-дың математикалық лингвистика зертханасында (НИИММ) ағылшынша-орысша (англо-русский) аударма жүйесі тәжірибелік тексеру дәрежесіне дейін жеткізілді [Лейкина және т.б., 1966]. Бұл жүйеде талдау мен жинақтауды (анализ и синтез) бір-біріне тәуелсіз қарастыру жағы есепке алынды. Ал өзгерту (преобразование) мен түсініктеме берудің (интерпретация) міндеттері – бірізділік қалыпқа келтіру. Талдау (анализ) әдісі фразалық және көптік сипатта іске асты. Тәуелділік грамматикасы моделінің шеңберінде тұжырымдалған лингвистикалық мәліметтер, оларды өңдеу шараларынан (алгоритмдерден) бөлек түрде қарастырылады. Алгоритм жұмысының нәтижелігі синтаксистік талдауды дұрыс жүргізу шарттарын қамтамасыз ету арқылы іске асты. Бұл шарттарға жататындар: жобалау (проективность), тарамдалу (древесность), байланыстың міндеттілігі, ортақтығы және т.б.

Аталған зертханада көп ұзамай байланыстың тұйықталу сипатын артық санау мен реттілікті есепке алатын орыс тілінің синтаксистік жүйесін құру және оны тәжірибелік сынаудан өткізу жұмысы басталады [Кремнева және т.б., 1976].

Ю.Д.Апресянның жетекшілігімен аса күшті лингвистикалық жабдықтауға ие французша-орысша аударма жүйесі құрылды («Информэлектро») [Синтаксисческий компонент..., 1981]. Аударма үстіңгі (поверхностный) синтаксистік құрылым деңгейінде жүргізіледі. «Информэлектро» (ИНФ-ЭЛ) негізінде іске асады:

1) аударуға тиісті тілдердің грамматикаларының баяндалуы «мән-мағына – мәтін» («смысл-текст») атты біріңғай моделімен келісу негізінде құрылды;

2) талдау мен жинақтау (анализ и синтез) ережелерінің бір-біріне ауысу, олардағы ақпараттың жеткілікті болуына негізделеді;

3) грамматикалық және сөздіктік баяндаулардың (описаний) алгоритмдерден, яғни программалардан да қашықтығы (аударманың өз ережелерінен басқа);

4) аударуға қатысты тілдердің баяндауларының (описания) тәуелсіздігі;

5) әр түрлі пәндік салаларға бейімделу тұрғысынан әмбебаптық;

6) лингвистикалық білімді көрсету үшін формалды аппаратты (модель және тіл) құру.

Бірінші кезекші ИНФ-ЭЛ жүйесі «ИРИС-50» ЭЕМ-де іске асырылды.

М.Тореза атындағы МГПИЯ-да АРМАС атты МА жүйесі үшін көпаспектілі ағылшынша-орысша сөздік құрастырыла бастады. Оның мақалалары сөздік бірілігінің әр алуан лингвистикалық сипаттамаларын қамтыды: морфологиялық, синтаксистік, семантикалық, лексикалық қызметі жайлы мәліметтер, басқару моделі жайлы аударма баламалары жайлы және т.б. [Шаляпина, 1974].

Бұл сөздік АРАП атты ағылшынша-орысша автоамтты аударма жүйесіның лингвистикалық жабдықталуының маңызды бөлігі ретінде енеді (жетекші-З.М.Шаляпина). Осы аталған жүйе үшін зерттеліп, қабылданған принцип (ұстаным) ЯРАП атты японша-орысша автоамтты аударма жүйесі үшін де қолданылды (жетекші-З.М.Шаляпина).

Бұл жерлер үшін талдаудың синтаксистік деңгейінің біреуінің болуы тән – семантика-синтаксистік. Тілдік бірліктердің семантикалық тіркесімен жазу мақсатында арнаулы тіл құрастырылған. Барлық негізгі лингвистикалық ақпарат біртиптік сөздік мақалалары түрінде көрініс тапқан. Оларда сөздік бірліктері грамматикалық, лексикалық, тіркесімдік және басқа да әр фразаны аударуға жеткілікті белгілер түрінде көрсетілген.

Бүкілодақтық аударма орталығы жұмысының екінші кезеңінде бірнеше аударма жүйелері құрастырылды. Осылардың ішінен тек екеуін ғана атап өтейік: Н.Н.Леонтьева жетекшілік еткен ФРАП атты французша-орысша автоматты аударма жүйесі [Ленонтьева, 1987] және Ю.Н.Марчук жетекшілік еткен АМПАР атты ағылшынша-орысша автоамтты жүйесі [Марчук, 1983].

ФРАП жүйесі аудармашының әрекетін модельдеуге арналған қызық талпыныс деуге болады. Аудармашы үшін аударылатын мәтін тілі ана тілі емес деп саналып, мәтінде белгісіз тілдік элементтер кездесетіндігі де есепке алынған. Ғылыми тұрғыдағы мұндай зерттеудің мақсаты – таза грамматикалық тәсілдің «күшін» анықтау. Осыған байланысты аударма процесі көптеген топтамаларға (блоктарға) бөлінеді де, ал олардың әрбіреуі аударуға қажетті фраза талдау деңгейінің белгілі сатысына сәйкес келеді. Тек формалды-грамматикалық құралдардың барлық мүмкіндіктерін қолданып болғаннан кейін ғана аударма жасау шарлары талдаудың семантикалық деңгейін қосуды ұсынады. Мұндай деңгей де грамматикалық сияқты сан жағынан бірнешеу болу қажеттігі алдын ала ескерілген. ФРАП жүйесіндегі аударманың концепциясы жүйенің синтаксистік компонентінің «күшін» дәлелдейді. Жүйенің ерекшелігі – дамыған семантикалық аппарат Бірінші синтаксистік топтаманың жұмысынан кейін ғана семантикалық аппаратқа

бетбұрыс жасау қажеттілігі алдын ала ескертілген. Жүйенің әр деңгейіндегі кішігірім жүйелері де тәжірибелік сынақтан өтетіндегі қарастырылған.

Грамматикалық және дамыған лексикалық талдауларға бағышталған АМПАР жүйесі де көп деңгейлі: сөз айналымдары, бірмағыналы және көпмағыналы сөздер жеке топтамалар арқылы өңделеді. Сондықтан тек сөйлем құрылымдары өзгеріске ұшырады. Аталған жүйе іс жүзінде (практикада) қолданыс тапты.

1970 жылдан бастап Р.Г.Пиотровский жетекшілік ететін Бүкілодақтық «Статистика речи» тобы МА жүйесі бойынша зерттеулердің негізінде қазіргі кезде де – актуалды ипеативті-стратификациялық тәсіл қолданылады. Мұндай әдіс бойынша аударуға тиісті мәтінді қайшылықсыз сипаттау (баяндау) үшін, біріншіден ол кезең-кезең бойынша (по этапно) қарастырылуы қажет, екіншіден зерттеудің соңғы нәтиже алынуы міндетті емес. А.И.Герцен ЛГПИ-дің инженерлік лингвистика зретханасында 1975 жылдан құрастырыла бастаған СИЛОД атты аударма жүйесі көптілді жүйе ретінде зерттелді. Оның сөздіктер менморфологиялық жабдықталуы аса күшті деңгейде және трансферді қолданатын грамматикасы бар. Осы кезеңдерде ағылшынша-орысша және французша-орысша МА-ның версиялары құрастырылып, тәжірибелік сынақтан өтті. Осының нәтижесінде жекеленген синтаксистік конструкциялардың құрылымы мен мағынасын тануға мүмкіндік жасайтын талдаудың фреймалық және тезаурустық негіздері қаланды [Пиотровский, 1979; Чижаковский, Беляева, 1983]. Аталған жүйеде МА-ның бірінші кезеңінде ұсынылған «фулькрумдар» әдісінң дамыған түрі ретінде МА-ның алгоритмдерін деңгейлік тұрғыда құру ескерілген.

МА жүйелерін құрастырудың екінші буының қысқаша қарастыра отырып, олардың негізі сипаттамалары мен ондай жүйелердің мынадай жалпы ерекшеліктерін атап кетуге болады [Кулагина, 1979; Hutchins, 1986]:

1) көпдеңгейлі талдау және синтаксистік деңгейдің толық дәрежеде берілуі;

2) талдау нәтижелерінің көпварианттылығы. Бұл жағдайда көптеген мүмкіндіктердің ішінен дұрыс біреуін таңдап алудың ықтималдығын арттырады.

3) лингвистикалық мәліметтерді сақтайтын қатал алгоритмнен барынша «жұмсақ» жүйеге ауысу. Соңғы жүйе бойынша барлық лингвистикалық мәліметтер, оларды өңдеу механизмінен, яғни алгоритмнен, бөлектелген [Лейкина және т.б. 1986; Кулагина, 1979].

Жоғарыда атлаған жүйелердің сипаттамалары мен ерекшеліктері МА жұмысын ұйымдастыруға оң әсерін тигізді. Тілшілер мен математиктердің жүйеніңерекше аспектілерін қарқынды зерттеуге бағыт-бағдар беріп, ынталандыруға себепші болды.Сонымен бірге аталған жайт тілдің формалды моделін құруды зерттеу мен оны дұрыс қолдану мүмкіндіктерін іздестіру қажет етті. Ал табиғи тілді формалды сипаттау грамматиканы бірізділікке салуға мүмкіндік туғызды. Сондықтан МА-ның даму кезеңінің екінші сатысында әр түрлі тілдерді сипаттау үшін формалды модельдер қолданылды

және де тілдік деректерді барынша барабар (адекватно) баяндауға мүмкіндігі мол жаңа модельдерді құру қажет болды.

Сонымен, МА-ның дамуының екінші кезеңі бірінші буынға қатысты практикалық жүйелердің дамуымен, тәжірибе жинақтау мен оны іске қосумен, сөздік базасының кең түрде өріс алуымен және тәжірибелік тұрғыда болса да барынша жабдықталған теориялық жүйелерді құрумен сипатталады.

Қол жеткен практикалық және теориялық нәтижелер МА-ның дамуының үшінші кезеңіне көшуге болатындығының айғағы болды. Бірақ келесі сатыға өту жолы алдын ала бірнеше экстралингвисикалық деректерді анықтап алуды қажет етті.

Біріншіден, 70-ші жылдардың соңында ақпарат ағымы бұрын-соңды болмаған қарқынды дамиды. Мысалы, америкалық және канадалық корпорациялар бір жылда 2,5 млрд. Сөзге дейін аударып отырған. Егер АҚШ-та 1966 ж. Аударма жұмысына шығарған біржылдық шығына 22 млн. доллар болса, 1981 ж. Ол шағын 200 млн. фолларға дейін өсті. Кеңес одағында штаттағы аулармашылардың айлығын санамағанда, тек заказ бойынша орындалатын аударма жұмысының өзіне орасан үлкен қаржы жұмсалады екен. Бірақ ақпарат ағымының көлемі (ғылыми-техникалық мәтіндер, патенттер, құжаттар мен арнаулы қағаздар) соншалық үлкен (зор) болуынан, оларды тек аудармашылар күшімен аударып шығу мүмкін емес. Сол себепті, техникалық аудармашылардың еңбегін аудару процедурасін автоматтандыру негізінде жеңілдету, жетілдіру қазіргі кезде маңызды мәселеге айналып отыр.

Екіншіден, спеллер типі өндірістік лингвитсикалық процессорларды құрастыу мен оларды іс жүзінде қолдану, сол сияқты сөздерді өңдеу жүйесін (word processors), күшті терминологиялық банктерді пайдалану – «лингвистикалық» автоматтанған жүйелерді қолданудан практикалық нәтижелерге ие болу мүмкіндігінің кепілі.

Үшіншіден, МА-ның жаңа кезеңіне өтуге ерекше жасалған алғы шарт есептеу техникасының дамуына жағдай жасады. – дисплейлердің пайда болуы, дербес компьютерлердің мен сканердің өмірге келуі, дербес ЭЕМ-ның күш-қуатының өсуі және т.б. Міне, осының бәрі теоретика-лингвистикалық зерттеулерге көзқарасты басқаша өзгеріп, оны практикалық жазықтыққа көтерді.

МА жүйелерін құрудың бірінші және екінші кезеңдерін сөз еткенде, біз нақты лингвистикалық ұжымдарға айырықша назар аударды деуге болады. Қазіргі кезеңді сипаттау үшін МА-ның стратегиясын таңдауға және МА-ның практикалық жүйелерін тиімді пайдалануға байланысты негізгі теоретикалық және техникалық мәселелерді анықтауға әрекет жасалық. Әрине, МА-ның мұндай негізгі мәселелерін талқылау МА жүйесінің болашақтағы даму жағдаятын да сөз етуді қажет етеді. Бұл жөнінен МА-ның дербес практикалық құндылығы бар екендігін және оның мәтіндерді автоамтты өңдеудің (МАӨ) көпфункционалды жүйесінің құрамына енетінін ескеру қажет болады.

Бірінші мәселе мынадай мүмкін болатын үш түрлі теориялық стратегияларды талқыға салуға саяды:

- тікелей бинарлық МА;
- трансфер арқылы аудару;
- аралық тіл (АТ) арқылы аудару, сонымен бірге олардың грамматикалық мүмкіндіктерінің ерекшеліктерін іске асыру арқылы.

Тікелей бинарлық аударма МА-ның «ескі» жүйелерінде орын алса да сол қалпына МА-ның көптеген практикалық жүйелерінің негізгі стратегиясына айналып отыр. Мұндай тәсіл нақты тілдік жұпқа бағышталады (GAT^A, SYSTRAN жүйелері және олардың дереваттары Бүкілодақтық аударма оратығында зерттелген АРМАС жүйесі). Бинарлық аударма әрекет кезінде сегменттердің шекарасын анықтау мен омонимияны мәнмәтіндік тұрғыда ажырату негізінде сөйлемдерге қарапайым өзгертулер (трансформация) жүргізілді.

Ертеректе құрылған МА жүйесін құрастырутәжірибесі мен оны пайдалану, мысалы, Джорджтаун университетінің GAT жүйесі мен оның түрлі варианттарын алатын болсақ, бұл жүйелердің тиімділігі көптеген жылдарбойы әрекет етуінің негізінде жинақталған аса күшті сөздіктер қоры мен арнайы редакциялау мүмкіндіктері арқасында деуге болады. Бірақ жеке лексика-морфологиялық және семантика-синтаксистік ішкі құрылымдар (подструктуры) негізінде іске асатын тікелей бинарлық сәйкестендіру жүйенің лингвистикалық және программалық жабдықталуын тек нақты жұп тілдердің ерекшеліктеріне ғана емес, сонымен бірге ішкі тілдердің (подъязык) айырым-белгілеріне қатаң түрде бағыштайды. Жұп тілдердің біреуі басқа тілге өзгертілсе, со бойында лингвистикалық, ал кейде программалық жабдықтауларды қайта құру қажет болады.

Трансформациялық МА (трансфер арқылы аударма) сондай-ақ МА дамуының екінші кезеңінде қолданылған. Трансфер пайдаланатын бұл стратегияны қолдану түрлі тілдердің мәтіндеріне тікелей қатысты жазбаларды жасаудың мүмкін еместігі жөнінде теориялық болжамдарға негізделеді. Бұл байланыс әртүрлі деңгейдегі жалғас жұмсалымдар жолымен анықталуы мүмкін: лексикалық, морфологиялық, синтаксистік құрылымдар деңгейінде (конфигурациялық және реляциондық синтаксис терминдерімен) және синтаксис деңгейінде сипаттау.Түрлі деңгейдегі талдаулар кезінде кіретін сөйлемнің құрылысын сипаттайтын трансферді пайдалану бағыныңқы таралымнан базалық грамматикалық бірліктер ретіндегі астырт басты құрылымға біртіндеп өтуді көздейді және одан әрі шығатын бағыныңқы таралымға өтеді. Мұндай кезде тек белгілі бағыныңқы таралымға қатысты қайта құру әрекеті сөйлемнің толық түсінігінің моделін құрастырмай-ақ МА жасауға мүмкіндік береді. Бұл мәселені МА кезінде шешу міндетті емес, өйткені астырт семантикалық-синтаксистік құрылымдарға түсініктеме беру тұтынушының үлесіне қалдырылады, яғни ЭЕМ-нен шыққан аудармамен әрі қарай жұмыс істейтін маманның үлесіне тиеді. Трансферді қолданудың мұндай тәсілі көптеген шетел ұжымдарымен бірге МА жүйесін қолданатын бірқатар ресейлік ұжымдарға да тән.

Бұл жүйелерде аударманың негізгі үш сатысын бөліп айтуға болады: талдау (анализ), трансфер және жинақтау (синтаз). Әрбір саты жүйелі субпроцестер түрінде орындалады. Мәселен, жоғарыда аталған SUSY атты жүйеде талдау әрекеті 8 субпроцестерді қамтиды. Нәтижесінде талдауға қатысты кіретін сөйлемнің таралымы (дерево) құрастырылады және шығатын сөйлемнің таралымы жайлы мәлімет 8 субпроцессордан тұратын трансфер түрінде көрініс табады. Талдау кезеңінде бұл таралым үш субпроцессорлардың өзара әрекеттесуінің нәтижесінде аралық тілдегі (АТ) сөйлем түріне қайта құрылады. Ұқсас тәсілдер трансфермен әрекет ететін басқа да жүйелерде қолданылады: CADA, GETA, TAUM-METEO, япон жүйесі, EUROTRA жүйесі, Ресейде СИЛОД (MULTIS) жүйесі, А.И.Герцен атындағы РГПУ-да зерттелген ФРАП, АРАП, ЭТАП, ЭТАП-2, STYLUS және т.б.

Европалық экономикалық бірлестігінің айырықша комиссиясы жанында 1982 жылы МА-ның дамуының үшінші кезеңінде ұсынылған EUROTRA жүйесінің негізі қаланды. Аталған жүйенің тілдері: ағылшынша, французша, немісше, итальянша, испанша, португалша, голландша, датша, жаңагрекше. Бұл жоба БЭЖ (ЕЭС) елдерінің тоғыз ресми тілдері үшін МА жүйесін құруды және 72 жұп тілдер бойынша аударма жасауды қарастырады.

Аталған тілдердің қайсысының болса да аудармасын жасауға қажетті талдауда интерфейстік көрінісі деп аталатын синтаксистік қатынастардың семантикалық түсінігі беріледі. Мәтін – семантикалық бағыныңқылықтарының түзілістері арқылы көрініс табады. Мысалы, олар астырт септігі типті болуы мүмкін (типа глубинных падежей). Атап кетуге тұрарлықтай жайт, ол бір тілдің интерфейстік құрылымы басқа тілдегі осындай құрылыммен сәйкес келмейді, бірақ трансферді қолдануға мүмкіндік туғызады. Сонымен, өз тілінің аясындағы әрбір маманға талдау мен жинақтаудан басқа трансфердің 8 процедурасын құра білуі қажет.

Жобаның мақсаты өндірістік МА жүйесінің алдын алу үлгісін жасау. EUROTRA жүйесінде іске асатын трансфердің ерекшелігі, оның қадамдап әрекет етуі. Әрбір қадам жасаудан бұрын, алдыңғы қадамдағы әрекеттердің бұрыс нәтижелері «тазартылады».

Трансфердіпайдаланатын жүйелерге ЭТАП-2 атты жүйе де жатады [Апресян и д.р. 1989]. Бұл жүйе ағылшынша-орысша МА жасайды, идеологиялық тұрғыдан ИНФ-ЭЛ жүйесінің жетілген түрі және айрықша сөздіктермен жабдықталған – нақты пәндік салаға бағышталмаған ұластыруға негізделген автоматты сөздік (АС). Зерттеушілер мұндай «тәуелсіздік» мүмкіндігін тілдік модельдің жан-жақты зерттелумен байланыстырады.

Трансфердің барлық ережелері жүйе ішінде жалпылық, жекелік (трафареттік) және сөздіктік деп үшке бөлініп қарастырылады. Жүйеде 45 қарапайым (элементарлық) предикаттар және 5 синтаксистік қатынастар орын алған сүзбеден өткізу тетігі (механизмі) мен жоғары бағалау ережелері синтаксистік болжамдардың ішінен біреуіне ғана артықшылық беруге мүмкіндік туғызады.

Трансферді пайдалану талдаудың әр түрлі деңгейінде кіретін сөйлемнің құрылымын сипаттайтын ішкі тармақтардан (поддеревья) базалық грамматика берілісін қолданатын астырт басты құрылымға жүйелі түрде көшуді және одан әрі шығатын ішкі таралымдарға көшу қарастырылған. Осындай жағдайда белгілі ішкі тармақтарға әрекеттері сөйлем түсінігін модельдемей-ақ МА жасауға мүмкіндік жасайды. Ұғынуды модельдеу міндетті МА мәселелерінің аумағына жатпайды, өйткені МА мәселесіндегі астырт түсініктеме алу адам үлесіне, яғни нақты пәндік саладағы маманға жүктелетіні жоғарыда да аталды.

Трансфер арқылы МА жүйелерін іске қосатын мұндай тәсіл МА дамуының екінші кезеңінде пайда болған Силод жүйесінде де қолданыс тапты. Қазіргі кезде SILOD-MULTIS көптілді жүйесі (персоналды) компьютерлерде аударма жұмысын атқарады. МА-ның барлық процесі лексика-морфологиялық және семантика-синтаксистік субпроцестердің композициясы ретінде SILOD-MULTIS жүйесінде модельденеді. Лексика-морфологиялық талдау кезінде мәтіндер мен сөздік бірліктерін теңестіру (идентификация) жұмысы жүргізіледі және автоматты сөздік топтамасынан сөздікке қатысты ақпараттар шығарылады. Мұндай талдаудың нәтижесі ретінде мәтіннің кесте түріндегі көрінісін және канондық пішіндегі аударма баламалары (переводные эквиваленты) жайлы ақпаратты айтуға болады. Бұл кестеде әрбір сөзтұлғаға оның лексика-семантикалық сипаттамасы сәйкестікке қойылады. Сәйкестіктің дәлдігі конверсиялық және/немесе омонимдік септелуге (подежная омонимия) дейін, функционалдық, басты (ролевой) және семантикалық сипатына дейін анықталады. Мәтінді талдаудың барлық деңгейлеріндегі көрінісі (берілісі, представление) сияқты аталған көріністер әр түрлі жағдайда қолданылуы мүмкін (МА, аннотациялауда, рефераттауда және т.б.). Егер МА жайлы айтатын болсақ, онда талдаудың бұл сатысында алынатын нәтиже – лексика-морфологиялық және фразеологиялық тұрғыда жолма-жол (подстрочник) аударылған мәтін.

Семантикалық-синтаксистік өңдеудің нәтижесінде құрылымды тіларалық деңгейінде қайта құру орындалады. Ол үшін лексика-морфологиялық талдау сатысындағы барлық ақпарат пайдаланылады, одан кейін грамматикалық және семантикалық топтамалар іске қосылады. Бұл трансфер иерархиялық бөліктері (компоненттері) мәтін ішінен алынатын тік (вертикально) қисындасқан (сопряженный) жүйешіктердің (подсистемы) жиынтығы (комплекс) ретінде модельденеді.

Семантика-синтаксистік талдау сөйлем құрылымының әр түрлі деңгейлерінде жүйелі түрде іске асады: топтар деңгейі, функционалды сегменттердің деңгейі және ең соңында, сөйлем деңгейінде [Беляева, 1986; Апресян және т.б., 1989, Соколова, 1990]. Топтар деңгейінде топтардың шекаралары мен ЛБ-нің өзара байланысы белгіленеді. Келесі деңгейде бөлініп алынған тізбектер (последовательности) сөйлемнің тұрлаулы мүшелеріне сәйкес келетін функционалды сегменттерге (топтарға) біріктіріледі. Бөлініп алынған сегменттерде мәтінді кестелік көрініске модификациялау деңгейінде қайта құру әрекеті орындалады. Функционалды

сегменттер (топтар) деңгейінде алынған нәтиже ретінде шектеулі грамматика бойынша жасалған МА болып табылады. Толық түрдегі семантика-синтаксистік талдау сөйлем деңгейінде ғана іске асады.

Талдау мен трансфер процедураларын жүзеге асырғанда синтаксистік сипаттау маңызды рөл атқарады, өйткені тек осының негізінде сөйлемнің мүмкін болатын фрейімі танылады. Мұндай таным әрекетінде актанттар жиыны, меңгеру моделінің мүмкіндігі жайлы және т.б. ақпараттарды бойында сақтайтын етістіктің лексика-синтаксистік сипатталуы негізгі рөл атқарады.

Осы әдістемеге сәйкес SILOD-MULTIS МА жүйесі автоматты сөздіктер мен грамматикалардың жинағы ретінде жүзеге асады. Жүйенің негізгі пішін үйлесімділігінде үндіеуропа және басқа құрылымдық тілдермен әрекет ету мүмкіндіктері алдын-ала қарастырылған. Мұндай мүмкіндік барлық «кіретін» тілдердің сөзтізбелерінің әмбебаптық құрылымына және «шығатын» орыс тілінің біріңғай құрылымына негізделеді.

Нақты әрекеттегі МА жүйелерінің трансферді пайдалану дәл осы стратегияны таңдаудың мақсатқа сай екендігін дәлелдей түседі. Сонымен бірге жасанды интеллек жүйесінің даму деңгейі және есептеу техникасының жаңа мүмкіндіктері МА-ның үшінші стратегиясына қызығушылықты жаңаша түрде жандыра түсті.

Аралық тілді (АТ) қолдануға қатысты аударма стратегиясы МА прагматикасына білім базасын құруда МА мен аның идеологиясының әсерін бейнелейді. МА-ның дему сатысының 1-ші кезеңінде қолданылған АТ тәжірибе жүзінде ғана іске асқанымен, ол өз уақытынан айтарлықтай озды – оның пайда болуы МА-ның жаңа кезеңі деуге болады. Мысалы, АТ-ны қолдана отырып, Карнеги университетінде CMU жүйесі, Утрехте зерттелген DLT жүйесі, Японияда – JETR, АҚШ-та – SAM және TRANSLATOR жүйелері машиналық аударма жасауға қолданыс тапты.

Аралық тілді пайдаланатын МА жүйесі мен жасанды интеллект жүйесін бір-біріне қисындастыру сөйлем мағынасы немесе тұтас мәтіннің мазмұнын тануға және автоматты түрде оның мазмұнын сөйлеу тілінде баяндауды шығатын тілде туындату үшін күрделі білім базасын қолдану қажеттілігімен түсіндіріледі (анықталады).

Жасанды интеллект жүйесінің аралық деңгейі болып саналатын түсінік алудың (понимание) концептуалды деңгейін модельдеу бірнеше базалық құрастырушыларға негізделеді:

- ұғым түйіндер (узлы-понятия) мен байланыс типтері жайлы сөз болатын мақала түріндегі сөздіктер құрастыруды қажет ететін білімді танытатын тілдер;

- белгі – мағына принципі бойынша қайта құруды жүзеге асыратын жүйелі грамматикалар.

Қазіргі кезеңдегі аралық тілдің (АТ) негізі ретінде білімді танытатын тілдер саналады. Осыған қатысты талдаудың мақсаты «кіретін» сөйлем бойынша алынатын және білім базасының ақпаратымен толықтырылатын сөйлем мағынасын таныту. Аударманы туындау әрекетінде «кіретін» тілдің

құрылымын «шығатын» тілдің құрылымына сәйкестендіріп, қайта құру қажет болады. Әрине мұндай аударма нәтижесі шын мағынасындағы аудармадан гөрі берілген мәтіннің мазмұнын не сөйлемнің мағынасын қайталап айтып беру жағдайында көбірек ұқсайды.

Осындай жағдаятта білімді танытатын тілді аралық тілдің (АТ) өзі деуге болады. Мұндағы АТ «кіретін фразаның (сөйлемнің) семантикалық желісі мен «шығатын» фразаның ішкі беоілісінен қайта туындауына қатынасты семантикалық таңбалауды («разметка») жүзеге асырады. Осылайша қайта туындату (генерация) белгілі үлгі бойынша (шаблон, фрейм және т.б.) орындалады. Бұл әдістің жалпы-теориялық тартымдылығына қарамай, ол автоматтандырылған лексикондардың негізінде аса күшті семантикалық желі құру үшін лингвистикалық негіздемесі мен программалық құралдары толық жасамай тұрып, тек тәжірибелік деңгейде ғана қалады.

МА жүйелерінің кең түрде қолданыс табуы оның эксперименттік режимнен автоамттанған дәрежесіне, редакциялау ерекшелігіне мақсатқа сай келуіне қарай, аудармашының жұмыс орнына айналуын мүмкіндік тудырады [Machine Translation, 1987; Апресян және т.б., 1989].

МА жүйелерінің қазіргі сатыдағы даму ерекшеліктері қарастыра келе, оның әр кезеңге бөліп классификациялаудан (топтаудан) мойын тартқан дұрыс сияқты. Себебі МА-ға қатысты ондай топтаулар есептеу машиналарының (компьютерлердің) кезеңді өзгерістеріне және мүмкіндіктерінің артуына тікелей байланысты.

Қазіргі кездегі МА жүйелерінің зерттелуі мен олардың қолданыс табуынан мынаны байқауға болады. МА жүелерін талдау және олардың функционалдық критериялары бойынша салыстыру мынаған, саяды:

1. Жүйенің прагматикасы. МА жүйесін өндірістік немесе эксперименттік дәрежеде құруға байланысты оның негізгі сипаттамасы мен талдау әдістемесін таңдау түрлері ерекшеленеді. Бұл критерий сөздік базасының көлеміне қойылатын талаптарды анықтайды: өндірістік МА жүйесінде бір жұп тілдерге 20000 сөздік мақаладан кем болмау. Мысалы, Bravice – Weidner LOGOS, GAT, S’YSTRAN және т.б. жүйелерде сөздік мақала көлемі көрсетілген көлемнен кем емес. Бұл шартқа бағындайтын жалғыз жүе – TAUM-METEO. Оның сөздік мақала көлемі 1500 тең, өйткені бұл жүедегі «кіретін» мәтіндердің тақырыптары қатаң түрде шектелген. Тәжірибелік (эксперименттік) жүйелерде автоматты сөздіктер «төменгі» жағынан шектелмеген, яғни 200 сөздік мақаласы бар сөздіктер бойынша да ол жүе аударма жұмысын атқара алады.

Өндірістік МА жүйесін құрудың мақсаты анық – үзіліссіз процесс түрін ұйымдастырылған құжат ағымдарынан МА алу. МА-ның өндірістік жүйелері мынадай талаптарға сай келуі керек:

- Жүйе жұмысының тұрақтылығы (устойчивость). МА жүйесінен алынатын нәтиже сапасын ойдағыдай болмауына қарамай тұтынушы оны өз қажеттігіне пайдалана алатындай болуы керек. МА жүйесінен шығатын ақпараттың сапасының төмен болуы, көбінде жүйенің сөздіктермен

жабдықталуының толық болмауына байланысты. Сондықтан жүйеден шыққан кейбір сапасы төмен («дөрекі») аудармаларәрі қарай маман аудармашының қолынан қайта түзетіліп сапасы жақсартылады. Осыған байланысты SYSTRAN жүйесін пайдалану тәсілімен танысу қажет.

- Жүйенің тиражданyы (таратылуы). Аударма жүйесінің қолдану аясы кең болуы үшін оның программалық және лингвистикалық құралдары барынша қарапайым түрде болуы қажет. Сонымен жүйеде бірге, басқа тақырыптық аяға және тілдерге көшу әрекетін жеңілдетуге қажетті әмбебептық модельдер орын алуы керек.

- Жүйенің икемделінуі (адаптациялануы). МА жүйенің нақты тұтынушының мұқтаждығына және өңделетін материалдың ерекшелігіне қарай икемделу мүмкіндігі (құралы) болуы қажет.

- Уақытша параметрлердің оптималдығы (ауқымдылығы). МА жүйенің аударма жасау жылдамдығы уақыт бірлігі ішінде қабылданатын ақпарат көлеміне немесе жүйені тұтынушы ретіндегі редакторлардың, аудармашылардың, мамандардың жұмыс әрекетімен сәйкес келіп отыруы қажет.

- Тұтынушыға жайлылық (комфорттық). Жүйенің сервистік құралдары тұтынушының жұмысының жұмыс істеу әрекеті әрі пакеттік әрі диалогтік режимдері үшін аса қолайлы болуы шарт.

Эксперименттік жүйелерді құрудың мақсаты МА-ның даму жолдарын зерттеу, адамның ойлау-сөйлеу қызметін модельдеу болып табылады. Тек осы тұрғыда ғана МА мен жасанды интеллектің міндеттері бір-бірімен сәйкес келе алады. Эксперименттік жүйелер мен оларды құрастырушығ ғалымдар МА-ның көш бастар рөлін атқарады, ал олардың жетістіктері өндірістік жүйелерде қолданыс табуы мүмкін. Мысалы, Саарбрюкен университетінің SUSY атты эксперименттік жүйесін құру арқылы алынған нәтижелер бойынша MARIS атты өндірістік жүйе құрылуда және SUSANNAH атты аудармашының автоматтанған жұмыс орны зерттеліп, жасалуда.

2. Жүйенің автоматтану дәрежесі. Практикалық тұрғыдан қарастырғанда МА жүйелерінің барлығы да адамның араласуын қажет етеді. Бірақ бұл критерийді қолдану жүйені екі қарама-қарсы көзқараста топтауға мүмкіндік береді: редакциялауда қажет ететін МА-дан бастап, аудармашы-адамға көмек беретін жүеге дейін (аудармашының автоматтандырылған жұмыс орны – Бүкілодақтық аударма жасау орталығы зерттеп, құрастырған СПАС жүйесін қара).

Шынында да, бұл жүйелердің негізгі айырмашылығы адамды аударма жұмысыныа тарту статегиясы – жүйе арқылы алынған нәтижені редакциялау немесе машинасыз (дәстүрлі) аудару кезінде МА жүйесінің аса күшті сөздіктер мүмкіндігін пайдалану. Осыған қарай редакциялау этапы мен оның көлемі айырым табады. Мысалы, МА-ның өндірістік жүйелерінде аударылған материалдың 20% қамтитын соң редакциялау (постредактирование) тәсілі кең түрде қолданылады (ALPS, ENGSPAN, LOGOS, METAL, SPANAM жүйелерін салыстыр). Әсіресе бұл жайт мәтіндер жиынын пакеттік режимде өңдеуді қарастыратын жүйелерге тән. МА

нәтижесі, ең алдымен, әрі қарайғы өңдеулерді қажет ететін бастапқы аударма нәтиже екендігін ескере отырып, МА мәселесін зерттеушілер арнайы сервистік жүйені құрудың және практикалық МА жүйесінің нәтижелігін бағалау үшін ондай сервистердің деңгейлік дәрежесінде ескеру қажеттігін оңғаруда. Мұндай сервис, әдетте, редакциялау құралдарын, сөздік жүелерін интерактивті енгізу құралдарын (автоматтанған лексикондарды) және т.б. өз құрамында сақтайды.

Ақпарат ағымының жаңа толқыны МА жүйелерін күнделікті өмірде кең түрде қолдану қажеттігін айқындап отыр, яғни ол үшін жоғарыда аталған мәселелер толық түрде ескерілуі керек.

ӘДЕБИЕТ

Автоматический перевод. 1949-1963 гг.: Критико-бтблиографический справочник / Сост. И.А. Мельчук, Р.Д. Равич М., 1967, 516 с.

Автоматический перевод: Сб. статей / Под ред. И предисл. О.С.Кулагиной, И.А.Мельчука. М., 1971. 357 с.

Андреев Н.Д. Машинный перевод и проблема языка-посредника // Вопр. Языкозн. 1957 №5. С. 117-121.

Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др. Лингвистическое обеспечение системы «ЭТАП - 2». М., 1989. 295 с.

Бар-Хиллел И. Будущее машинного перевода // Науч. докл. высшей школы. Филол. науки. 1962. №4. С. 203-206.

Белопольская А.Р., Ковригин А.Б. Исходный этап независимого анализа немецкого простого распространенного предложения // Материалы по математической лингвистике и машинному переводу. Вып. 2. Л., 1963. С. 150-160.

Беляева Л.Н. Применение ЭВМ в лингвистических исследованиях и лингводидактике. Л., 1986. 84 с.

Воронин В.А. Система независимого конфигурационного анализа китайского текста при машинном переводе // Машинный перевод. М., 1961. С. 135-172.

Дрейзин Ф.А., Рашитов Р.С. Принцип синтаксического анализа татарской фразы // Машинный перевод. М., 1961. С. 295-303.

Кремнева Н.Д., Морева В.Ю., Ульянова Г.С. Принципы построения грамматики русского синтаксического анализа с учетом предпочтения грамматики русского синтаксического анализа с учетом предпочтения и порядка замыкания связей // Лингвистические проблемы функционального моделирования речевой деятельности. Вып. 3. Л., 1976. С. 139-144.

Кулагина О.С. Исследования по машинному переваоду. М.,1979. 319 с.

Кулагина О.С., Мельчук И.А. Машинный перевод с французского языка на русский // Вопр. языкозн. 1956. №5. С. 111-121.

Лейкина Б.М., Никитина Т.Н., Откупщикова М.И. и др. Система автоматического перевода, разрабатываемая в группе математической лингвистики ВЦ ЛГУ // Науч.-тех. информация. 1996. № 1. С. 40-50.

Леонтьева Н. Н. Система французко-русского перевода (ФРАП): лингвистические решения, состав, реализация // Машинный перевод и прикладная лингвистика: Проблемы создания системы автоматического перевода. Вып. 271. М., 1987. С. 6-26.

Марчук Ю.Н. 1) Проблемы машинного перевода. М., 1983; 2) Методы моделирование перевода. М., 1985. 199 с.

Машинный перевод: Сб. статей. М., 1957. 314 с.

Мельчук И. А. 1) Работы по машинному переводу в СССР // Вестн. АН СССР, 1959. № 2. С. 43-47; 2) Опыт теории лингвистических моделей «смысл – текст». М., 1974. 314 с.

Молошная Т. Н. Алгоритм машинного перевода с англиского языка на русский // Проблемы кибернетики. Вып 3. М., 1960. с. 209-272.

Откупщикова М. И. Синтаксис связного текста. Л., 1982. 103 с.

Панов Д. Ю. Автоматический перевод. 2-е изд. М., 1958. 72 с.

Переводная машина вТалине в 1924 г. // Сообщения по машинному переводу. Сб. 1. Таллин, 1962.С. 95-96.

Переводная машина П. П. Траянского: Сб. матер. о переводной машине для перевода с одного языка на другие, предложенной П. П. Троянским в1933 г. / Отв. ред. Д. Ю. Панов. М., 1959. 52 с.

Пиотровсий Р. Г. Инженерная лингвистика и теория языка. Л., 1979. 112 с.

Синтаксический компонент в системах машинного перевода // ВЦП. Сер. 2. Машинный перевод автоматизация информационных процессов: Обзорная информация. Вып. 5. М., 1981. 130 с.

Шаляпина З. М. Англо-русский многоаспектный автоматический словарь (АРМАС) // Машинный перевод и прикладная лингвистика. Вып. 17. М., 1974. С. 7-67.

Шумилина А. Л. Вопросы анализа личных местоимений третьего лица // лингвистические исследование по машинному переводу. Вып. 2. М., 1961. С. 132-140.

Цейтин Г С., Откупщикова М. И., Лейкина Б. М. и др. Система анализа текста с процедурным представлением словарной информации // Актуальные вопросы практической реализации систем автоматического перевода. Ч. 1. М., 1982. С. 205-231.

Чижаковсий В. А., Беляева Л. Н. Тезарурус в систамах автоматической переработки текста. Кишинев, 1983. 163 с.

Hutchins W. J. Transiation: Past, Present, Future. Harwood, 1986.

Machine Transiation: Theoretical and Methodoiogikai Issues // Ed. by S. Nirenburg. New York, 1987. 350 p.

ҚАЗАҚ ТІЛІНІҢ «ТІЛ – ҚАЗЫНА» АТТЫ ҰЛТТЫҚ КОМПЬЮТЕРЛІК ҚОРЫН ҚҰРУДЫҢ АЛҒЫШАРТТАРЫ

«Профессор С.М.Исаев: Қазақ тіл білімінің

теориялық және әдіснамалық мәселелері»
атты халықаралық ғылыми-теориялық
конференция материалдары. – Алматы, 2008.
9-10 маусым. – 12-19 бб.

1950 жылдар электронды-есептеуіш машиналардың (ЭЕМ) күнделікті өмірге қарқынды еніп, гуманитарлық ғылым саласындағы зерттеулерде қолданыла бастады. Нәтижесінде, осындай әдіс-тәсілдерге қатысты көптеген ғылыми мақалалар мен монографиялар жарық көрді.

Соңғы жылдары ғылым мен техниканың дамуында гуманитарлық ғылымдар басымдық танытып, (және мүмкіндіктері әлі де толық ескеріле қоймаған) әлеуеті жоғары екендігі анықталды. Бұл салада Экономикалық математикалық басқару әдістерімен қатар әлеуметтік және коммуникациялық мәселелер де үстемдік алып келеді. Тілдегі Іскери құжаттарды өңдеуде, бұқаралық коммуникацияда т.б. гуманитарлық ғылымдарының маңыздылығы байқалды.

Қолданбалы тіл білімінің белгілі маманы Д.А.Поспеловтың сөзімен айтқанда «естественный язык является самой мощной из известных нам сегодня моделирующих действительность систем». Сондықтан «білім қорын» қалыптастыруда компьютерге табиғи тілді «үйрету» көзделеді. Әрине өндірісті автоматтандыру заманындағы мұндай қажеттілік көпшіліктің қажеттілігінен туындап отырғаны белгілі.

Есептеуіш лингвистиканың ірі теоретигі, белгілі ғалым Д.Хейс 1973 жылдың өзінде-ақ тіл білімінің ішкі мүмкіндігі жайлы былай деп жазған болатын: «Лингвистика таит в себе громадные неиспользовавнные возможности для точных наук, и пренебрегать ими далее недопустимо. Возможно, спустя тысячелетия, люди с недоумением будут думать, почему наука наших дней так часто пользовались математикой, оставляя без внимания простые, естественные лингвистические принципы».

Бұл өте орынды пікір, себебі «ЭЕМ-мен диалог жүргізу» тақырыбына арналған ғылыми жарияланымдар мен ғылыми мәжілістердің (семинарлар, симпозиумдер, конференциялар) күн санап көбейіп келеді.

Бүгінгі таңда − жаратылыстану ғылымынан гуманитарлық ғылымдарға әсері (және керісінше) күшейіп өзара сабақтасуда. Мәселен, қазіргі кезде әлеуметтік ғылымдағы көптеген тұжырымдар әдіс-тәсілдер жаратылыстану мен техника саласындағы ғылымдарға күшті әсер етуде. Ықтималдық теориясы, ойындар теориясы және ақпараттар теориясы, бір-біріне ықпал ету де. Мысалы, кибернетика саласы ғылыми ойлау стилінің дамуына елеулі үлес қосты және оның мазмұндық сипаты әлеуметтік ғылымдармен тығыз қатынаста болуына байланысты, әдіснамасы да күннен-күнге байып келеді. Сол сияқты, әлеуметтік және философиялық озық ойлар арқылы дамыған мақсаттылықтың, басқарудың, жүйелі талдаудың жалпы идеяларлары кибернетика саласына әсер етуде. Жасанды интеллектіге қатысты зерттеулер де кибернетика мәселелерінің гуманитарлық ғылым идеяларымен сабақтас екенін көрсетуі.

Гуманитарлық ғылымдардың математика саласына әсері бұрыннан белгілі. Тілдерді зерттеу де, әлеуметтік факторлар ескеріліп отырады. Мысалы, «Евгений Онегин» романы мәтінін зерттеу негізінде математикада «Марков тізбегі теориясы» атты заңдылық ашылды, әдеби ескерткіштерді зерттеу арқылы белгілі ғалым Юл статистикалық үлестірілулер күрделілігінің құпиясын ашты. Сол сияқты Колмогоровтың өлең құрылымын зерттеуі, Заденың «Теория размытых множеств» деп аталатан еңбегі – тіл білімінің математикаға әсерінің нәтижесі. Өткен ғасырдың 30-40 жылдардағы құрылымдық бағыттағы зерттеулердің жалғасы ретіндегі Н.Хомскийдің лингвистикалық ізденістерінің ықпалынан «математикалық лингвистика» саласы өмірге келді.

Тіл ғылымының әсерін қолданбалы кибернетика саласынан да байқауға болады. Нәтижесінде, соңғы кездері «автоматтанған жүйелерді лингвистикалық қамтамасыздандыру» деп аталатын ұғым қалыптасты. Қазіргі физика мен математикалық логикаға салаларына гуманитралық әдіс-тәсілдерінің енуіне байланысты Бордың «үстеме ұстаным», «модалдық-логиканың қалыптасуы» деп аталатын принциптері пайда болды. Соңғы кезде ғана жаратылыстану және математика саласындағы ғалымдар үшін ақпараттық ортада әрекет етудің маңыздылығы түсінікті бола бастады. Ақпарат ағымын басқару ұғымы табиғи ортадағы экологияны басқару ісімен ұқсас деуге әбден болады.

Гуманитарлық ғылымдарда мамандандырылған мәселеге бағытталған автоматтанған жүйелерді кең түрде қолдану қажет. Мысалы, қазақ тіл білімінде автоматтанған лексикографиялық жүйе арқылы мәтіндермен, сөздікпен жұмыс істеуге (іздеу, өзгерту, толықтыру және т.б.) болады. Сол сияқты автоматтандырылған ақпараттық жүйеде құжаттық дереккөздерді (заң шығаратын актілер, ғылыми-техникалық жарияланымдар, әдеби шығарламалар және т.б.) жинақталып, өңделеді. Сауалнама негізінде жинақтап қорытындыланған автоматтанған жүйелер әлеуметтік, тарихи және т.б. зерттеулер жүргізуге ыңғайлы. Әсіресе олар автоматтанған редакциялық-баспа жүйесіндегі деректерді, фактілер мен библиографиялық мәліметтерді тексеруде, қолжазба мәтіндерді автоматты түрде баспаға дайындауда қолданылып келеді. Тәжірибелерді автоматты басқару жүйелері де психология мен тәжірибелік фонетика және т.б. ғылым салаларында қолданыс тауып келеді.

Гуманитарлық ғылымдарда, әсіресе қазақ тіл білімінде ғылыми-зерттеу жұмыстарын жүргізу үшін қажетті де маңызды құрал ретінде жиналған мәліметтерді (деректер, дәйектеме, дереккөздер және т.б.) бейнелейтін картотекалық қордың маңызы ерекше. Сондықтан тілші-ғалым үшін компьютер дегеніміз – жадында сақталған өте зор көлемдегі картотекаларды бағдарлама көмегімен басқаруға арналған құрал. Компьютерді гуманитарлық зерттеулерде қолданудың ең алғашқы және ең негізгі сатысы – табиғи формадағы деректерді автоматты жинақтау болып табылады. Мысалы, тіл білімінде сөздіктер, грамматикалар және мәтіндер; әлеуметтануда – сауалнамалық нұсқалар мен сұхбаттар; экономикада – әлеуметтік-

экономикалық нысандарды сипаттайтын экономикалық көрсеткіштер жиынтығы; журналистикада – редакциялық дайындық кезеңіндегі анықтағыштар мен досье мәтіндері; психологияда – тәжірибелік мәліметтердің ағымы; тарихта – дереккөздердің формуляры мен ескеркіштер мәтіндері және т.б. Ондай мәліметтер зерттеушінің ұсынған үлгісі (моделі) бойынша формалардың логикалық, есептеуіш немесе статистикалық өңделеді.

Қазіргі кезде деректерді өңдеу математикадағы, ақпараттану мен тіл біліміндегі әдіс-тәсілдері зерттеудің екінші, ең қиын сатысына өтуді айтарлықтай жеңілдетті. Яғни проблемалық-бағыттық (белгілі бір мәселеге арналған) жүйелерде, бірінші сатыда жинақталған мәліметтерді (деректерді) шағын түрдегі қорытынды мен ғылыми болжам түріне келтіру ісі жетіліп келеді. Мысалы, ғылыми-техникалық ақпараттану саласында тиісті құжатты іздеп табуда белгілердің арнайы логикалық ұйымдасу сипаты негізге алынса, лингвистикада лексикографиялық және грамматикалық ақпаратты формалды түрге келтіру әдісі қолданылады, білім қорын лингвистикалық құралдар арқылы бейнелеу әдісі әлі де зерттелуде, экономикада бір көрсеткіш басқа көрсеткіштердің жиынтығы бойынша есептеп шығарылады және т.б.

Гуманитарлық ғылымдардың халықшаруашылық пен ғылыми-техникалық практикада қолданыс табуы нәтижесінде жетілді. Сонымен қатар жаратылыстану және техникалық ғылымдармен байланысы күшейіп келеді. Мұны әдетте гуманитарлық салаға қатысты білімнің математикаландырылуы ретінде сипатталады, шын мәнінде, гуманитарлық ғылымдар да өз тарапынан жаратылыстану мен техникалық ғылымдарға сүбелі үлес қосып келеді.

Сонымен, гуманитарлық ғылымдарда ЭЕМ-ді (компьютерді) қолдану қоғамдық, жаратылыстану және техникалық ғылым салаларын өзара жақындастыруға себепші болды, сонымен бірге автоматтандыру құралдарын жетілдіруге және оның қолдану аясын кеңейтуге жол ашты. Лингвистикалық зерттеулерді автоматтандыру қазақ тіл білімі үшін де әдіснамалық міндеттер жүктеуде. Мәселен, қолданбалы лингвистика дәстүрлі түсініктегі тіл мәдениетін арттыру, тілдерді үйрету, ғылыми-техникалық терминологияны қалыптастырумен ғана айналыспайды. Қазіргі кезде қазақ тіліндегі деректерді автоматты өңдеуге байланысты автоматты жүйені қазақ тілінің фактілерімен қамтамасыз етіп, компьютермен қазақ тілінде қарым-қатынас жасау мәселелері қолға алынуда. Қазақ тілін нормалауға бағытталған академиялық сөздіктер мен грамматика жарық көргенімен, сауаттылыққа үйрету, сөйлеу мәдениетін арттыруға қатысты жазылған еңбектер формасы, толықтығы, дәлдігі және тілдік материалдарды қамтуы жағынан компьютердің талабын толық қанағаттандыра алмай келеді. Тілдік деректерді өңдеуге арналған техникалық құралдар тәжірибе жүзінде шексіз мәліметтерді жинақтап қажетті үлгіде сақтауға мүмкіндік бере алады. Мәселен, автоматты жүйелерді жабдықтау үшін автоматтанған сөздіктер мен грамматика түрінде немесе қажеттілігін өтейтін әдеттегі сөздіктер мен грамматикалар түрінде де дайындауға болады.

Мұндай мүмкіндіктер қазақ тілшілерінің еңбек өнімділігін күрт артырады, сөздік дайындау ісін жетілдіреді, тездетеді, олардың түрлерін де көбейтуге болады. Алайда лингвистикалық зерттеулерді жаңа технологияға ауыстыру, дәстүрлі қағаз кәртішкі үйреншікті полиграфиялық форматтан бас тартатын болсақ, лингвистикалық еңбектерді дайындаудың жаңа технологиясына көшуіміз керек, яғни лингвистикалық құрылымдау теориясына негізделген лексикографияның жаңа теориясын зерттеу қажет.

Автоматтанған қазақ лексикографиясының іргелі мәселелерінің қатарында: тілді құрылымдық-семантикалық және функционалды-стилдік ең кіші бөлікке дейін жіктелген дереккөздердің саралап топтастыру бағытында зерттеулер жүргізу; сөздік мақаланың жалпылама құрылым-формасының сұранысына қатысты сөздіктерді автоматты түрде жобалау; мәтіндік және сөздік массивтерді индекстеу құралы ретінде автоматтанған морфемдік анықтағыштарды құрастыру мәселесін зерттеу т.б. атауға болады.

Әрине, бұл қазақ тіл білімі үшін принципті тұрғыда жаңа зерттеу нысаны болғанымен, үнді-еуропа тілдерінде ғалымдар бұл мәселелермен шұғылданып келеді. Оны ғылыми-техникалық революцияның қажеттігіне байланысты сапаның жаңа деңгейіне назар аудартудан туындаған жаңаша тұжырым ретінде қабылдау керек.

Бұл саладағы ізденістер тарихи, әлеуметтік және психолингвистикалық деректерді, диалектология мен лингвогеография, топонимика мен ономостика деректерін ұқыптылықпен тіркеп отыруды қажет етеді.

Жаңа мәселелер қазақ тіл білімінің аталған салаларына әрі жаңа, әрі жоғарғы деңгейдегі міндет жүктейді. Қазақ тілінің мәтін семантикасы мен лексикасы аясындағы ғылыми зерттеулердің және түсіндірме сөздіктердің маңызы арта түседі. Бұл мәселе «жасанды интеллектіні» құрастыру ісіне тікелей қатынасы бар, яғни «қазақша ойлай алатын» зияткерлі роботтарды басқаруда да аса маңызды. Қазақстан өндірісін қазақ тілінде толық автоматтандыруға жол ашылады, еңбек тиімділігін арттырып, әсіресе адамдардың денсаулығына зиян келтіретін аса қиын орталарда «қазақ роботтарын» пайдалануға мүмкіндік туады.

«Қазақ әдеби тілі» ұғымын анықтай түсу қажет. Өндіріс пен басқару тілін, ғылым тілі мен ғылыми-техникалық құжаттар тілін, халықшаруашылық салаларына тілдік қатынастардың барлық түрлерін енгізу қажет. Сонда ғана қазақ тілінің ұлтаралық және халықаралық қызметі күшейеді.

«Қазақ әдеби тілі» ұғымының кеңеюі, тілдің функционалды-стилистикалық саралануы жетіліп, жаңаша тұрғыда нақтыланып келеді. Сөздіктер мен грамматикаларда ескерілетін тілдік формалардың түрлері көбейіп келеді.

А.Байтұрсынұлы атындағы Тіл білімі институтында жаңадан құрастырылып жатқан 15 томдық академиялық «Қазақ әдеби тілінің сөздігінде» әдеби тілдің кеңейтілген ұғымы мен оның түрлі бағыттары ескерілуде. Яғни аталған сөздік нормативті-стилистикалық сипатта болуымен қатар, қазақ мәтінін автоматты өңдеудің де құралы болады деген сенімдеміз. Сонымен, қазақ әдеби тілінің сөздігі әрі өте бай, әрі формалды түрде айқын

ақпаратты және сөздердің синтаксистік және семантикалық қасиеттерін толық қамтуды мақсат ететін сөздік болады деген үміттеміз. Осындай ғылыми еңбектердің дамуы, қазақша мәтіндерді автоматты өңдеудің әдіс-тәсілдерін меңгеруге, әдебиеттану саласын автоматтандыруға да негіз болары сөзсіз. Нәтижесінде жаңа басылымдарды автоматты дайындау, ғылымдарға ақпараттық-анықтағыштық қызмет атқару, мәтіндерге қатысты анықтағыш құралдар дайындау, ғалымдардың зерттеу тәжірибесіне автоматтанған картотекаларды пайдалану және т.б. мәселелері қолға алынуда.

Алайда, гуманитарлық зерттеулерді компьютерлендіру оңай іс еместігін ескерте кету қажет. Оның нәтижесі деректерді автоматты өңдеуге қажетті аса зор мүмкіндіктерге ие техникалық құралдарда, жетілген математикалық және бағдарламалық материалдармен қамтамасыз етілуіне, тұтынушы-тілшінің дайындық дәрежесіне және есептеу ісінде қолданылатын лингвистикалық және дидактикалық құралдардың жеткілікті сипатта болуына байланысты.

Қазақ тілін зерттеушілер үшін арнайы компьютерлік орталық ашылып, оларда бағдарламалау тілдері, деректер қоры, лингвистикалық салада қолданылатын бағдарламалар пакеттері және т.б. жинақталуы қажет. Сонымен бірге бірнеше рет қолданылған мәтіндер, сөздіктер, ақпараттық материалдар, қоғамдық пікірге қатысты сұрақ-жауаптар және т.б. материалдар да қорлануы керек. Алғашында «тырнақтап» жинақталған архивтік қор болашақта тілшілердің, әлеуметтанушылардың, тарихшылардың зерттеулеріне аса құнды материал болуы мүмкін.

Мәтіндік қорды компьютердің жадында немесе сыртқы таспаларда сақтау технологиясын меңгеру редакциялық-баспа ісін автоматтандыру әрекетінде де аса маңызды. Мәселен, сөздіктерді, анықтағыштарды, энциклопедияларды, жинақтарды, оқулықтар мен оқу құралдарын, белгілі оқырман аясына бейімделген әдеби және ғылыми-көпшілік шығармаларын компьютер жадында сақтау технологиясын меңгеру өте маңызды.

Әрине, гуманитарлық зерттеулерге арналған компьютерлік орталық гуманитарлық білім иелерін, яғни маман емес тұтынушыларды дайындайтын оқыту орталығына айналғаны да жөн болар еді.

Жаппай автоматтандыру заманында қоғамның әрбір мүшесі белгілі дәрежеде, «компьютерді пайдаланушы» саналады. Сондықтан білімнің гуманитарлық саласына, әсіресе қазақ тіл біліміне, компьютер мүмкіндігін толығымен пайдалану егемен еліміздің дамуына қосқан зор үлес болар еді.

ӘДЕБИЕТ

1. *Андрющенко В.М.* Концепция и архитектура Машинного фонда русского языка. – М.: Наука, 1989. – 200 с.

2. *Глушков В.М.* Основы безбумажной информатики. – М.: Наука, 1982.

3. *Жұбанов А.Қ.* Тіліміздің компьютерлік қорын жасау керек //Ана тілі. – 2003. – 13 қараша (№ 45).

ЕСЕПТЕУІШ ЛЕКСИКОГРАФИЯ – ҚАЗАҚ ТІЛ БІЛІМІНІҢ ЖАҢА ҒЫЛЫМИ БАҒЫТЫ

Тілтаным (Языкознание). А.Байтұрсынұлы атындағы
Тіл білімі институты, Алматы. 2008, №2, - 58-63 бб.

Қазіргі кезеңдегі қазақ тілінің лексикография саласын автамоттандырудағы саясат табиғи тіл деректерін өңдеуге қатысты (редакциялық-баспа ісін қоса есептегенде) ой-пікірлер мен әдістердің бір жерге шоғырлану сипатымен ерекшеленеді. Бұл жағдай қазақ лексикография саласындағы сөздік құрастыруға қажетті картотека қорын жинақтау мен сөздікті жобалаудан бастап, баспа ісінде тілдік материалдарды автоматты теру, беттеу (верстка) мен таралымын белгілеуге дейін лексикографиялық қызметті кешенді түрде автоматтандырудың алғышарты деуге болады.

Мәселен, А.Байтұрсынұлы атындағы Тіл білімі институтында көп жылдардан бері жинақталған осындай картотекалық қордың жалпы көлемі шамамен алғанда 5 млн. мәнмәтінді құрайды. Осы қордың негізінде екі томдық, он томдық «Қазақ тілінің түсіндірме сөздігі», «Абай тілі сөздігі» және басқа да сөздіктер жарық көрді. Қазіргі кезде Институттың лексикограф-мамандары осы және басқа да тың картотекалық қордың негізінде 15 томдық «Қазақ әдеби тілінің сөздігін» құрастыру үстінде [1].

Қазақ тілінің сөз «қоржынындағы» мұндай көлемді картотекалық қорды толықтыру, түзету немесе ұзақ мезгілге сақтау мен пайдалану жұмыстары тілші-лексикографтар үшін оңай әрекет емес. Сондықтан да, заманға сай, қазақ лексикография саласын автоматтандыру (немесе компьютерлендіру) кешенді проблемасын шешудің бастамасы ретінде А.Байтұрсынұлы атындағы Тіл білімі институтында «ТІЛ – ҚАЗЫНА» атты қазақ тілінің компьютерлік картотекалық базасы (қоры) іске қосылды. Оның негізгі мақсаты – қағаз бетіндегі 5 млн.-дай картотекалық қорды сан мен сапа тұрғысынан қайта сұрыптап, компьютер жадына жазу, яғни «ТІЛ – ҚАЗЫНА» қорына енгізу, әрдайым ол қорды жаңа деректермен жаңартып отыру және пайдалануды барынша оңайлату.

Осы аталған мәселені шешудің тағы бір жолы – «Сөзнұсқағыш әліпби-жиілік сөздігін» жасау. Бұл сөздіктің басқа жиілік сөздіктерден негізгі айырмашылығы – баспадан шыққан шығарма мәтініндегі сөздердің кітап беттеріндегі орындарына нұсқау (мегзеу). Яғни ол сөзтұлғаның (сөздің) қай бетте және сол беттің қай жолында кездесетініне сілтеме жасау. Сонда зерттеушіге қажет деген сөзге қатысты мәнмәтінді сол сілтеме арқылы кітап беттерінен оңай тауып алуына мүмкіндік туады. Қазақ тіл білімінде осындай әдіс тұңғыш рет А.Байтұрсынұлы атындағы Тіл білімі институтында М.О.Әуезовтің 20 томдық шығармалар жинағының жиілік сөздіктерін алу жағдайында қолданылған болатын. Әрине мұндай сөздік көлемінің тым үлкен болуынан, арнайы сөздік ретінде баспадан шығарудың өз қиындықтары да жоқ емес. Бірақ мұндай «Сөзнұсқағыш әліпби-жиілік

сөздікпен» тілші қауымының автоматты сөздік ретінде пайдалану мүмкіндігі аса мол.

Лексикография саласын автоматтандырудың әлем бойынша қысқаша тарихын сөз ететін болсақ, бұл бағыт осыдан 55 жыл бұрын Роберт Буздың жұмыстарынан бастама алды. Белгілі ғалым әлемдік тәжірибеде тұңғыш рет әдеби ескерткіштерді зерттеуге қажетті, дәлірек айтқанда, Фома Аквинскийдің шығармаларына сөзнұсқағыш пен конкорданс құрастыруда есептеуіш-перфорациялық мәшинені қолданған болатын [2].

Ал 1956 жылдан мұндай жұмыстарға электронды-есептеуіш мәшинелер (ЭЕМ) қолданыла бастады. Осының нәтижесінде қарастыратын мәселелердің ауқымы да кеңейе түсті: әліпбилі кері, әліпбилі-жиілікті және жиілікті сөздіктер алу, картотекалар әзірлеу, тілдің лексикалық құрамын топтау мақсатында мәтіндер мен сөздіктерді өңдеу және неше түрлі статистикалық мәліметтер алу, семантикалық өрістерді құру сияқты жаңа технологияға сай келетін мәселелер [3; 4; 5; 6; 7]. Арнайы жабдықталған зертханаларда үлкен көлемді мәтіндер мен сөздіктерді ЭЕМ-де өңдеу әрекеттері Р.Буздың кезеңдерімен салыстырғанда 7-10 есе өскенін де айта кетуге болады.

Осындай лексикографиялық жұмыстарды автоматтандыру әрекеттері ұлттық тілдер, соның ішінде қазақ тілі қорларының автоматтанған жүйесін құру мәселесіне әкелгені де мәлім.

1969 жылы Қазақстан Ғылым академиясы Тіл білімі институтында «Статистическое и информационное изучение тюркских языков» атты Бүкілодақтық ғылыми жиынның өтуі қазақ тіл білімі үшін елеулі құбылыс болды деуге болады. Осы ғылыми форумның нәтижесі ретінде 1970 жылы Тіл білімі институтында статолингвистика саласының белгілі математик ғалымы, филология ғылымдарының докторы, профессор Қалдыбай Бектаевтың жетекшілік етуімен «Тіл статистикасы және автоматтандыру» атты ғылыми топ құрылды. Осы кезеңнен бастап қазақ тілін статистикалық зерттеу ісі мен лексикографиялық жұмыстарды автоматтандыру әрекеттері жүйелі сипатқа ие бола бастады. Негізгі статистикалық зерттеу құралы ретінде алғашында электронды-есептеуіш машиналар (ЭЕМ) пайдаланылып, олардың көмегімен қазақ тілінің әртүрлі стильдері бойынша әліпбилі-жиілік сөздік, жиілік сөздік және кері әліпбилі-жиілік сөздіктер түрлері құрастырылды. Олардың негізгілері: 1979 жылы «Ғылым» баспасынан жарық көрген М.О.Әуезовтің «Абай жолы» романының жиілік сөздігі, 1995 жылы сол баспадан басылып шыққан «М.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері». Бұл аталған жиілік сөздіктер ұлы жазушы М.О.Әуезовтің 100 жылдық мерейтойына арналған Қазақстандық статолингвистерінің арнайы сыйы еді. Жазушы тілінің лексикалық байлығын, автордың сөз қолданудағы стильдік тәсілінің өзіндік ерекшелігін, қазақ тілінің көркемдік және бейнелеуші мүмкіндіктерін меңгеру дәрежесін осы аталған лексикографиялық еңбектер бойынша бағалауға мүмкіндік бар. Кітап түрінде баспадан жарық көрмесе де, М.Әуезовтің 20 томдық шығармаларының әр томы бойынша ЭЕМ жадында сақталған «Сөзнұсқағыш әліпби жиілік сөздік» атты электронды сөздік базасы да арнайы қаралатын

лексикографиялық еңбек дейміз. Сонымен бірге Отандық түркітану ғылымы үшін ортағасырлық жазба ескерткіші «Кодекс Куманикус» мәтіні тұңғыш рет ЭЕМ жадына енгізіліп, ол бойынша автоматты түрде «Куманша-қазақша жиілік сөздігінің» алынуы және оның «Ғылым» баспасынан жарық көруі қазақ тілі тарихы үшін елеулі оқиға.

Аталған мәселенің қысқаша тарихын жалғастыра түссек, Францияның Нансий университетінде XVI-XX ғг. француз тілінің автоматтандырылған сөздік картотекасының толық нұсқасы ақпарат сақтағыш магниттік таспада орналасқанын айтуға болады. Оның әдеттегі сөздік картотекасынан айырмашылығы тек техникалық орындалуында ғана емес, сонымен бірге құрылымдық жағында да болып отыр. Мысалы, әрбір таңдалып алынған үлкен көлемдегі мәтіндер жиынын (шамамен алғанда 100 мың сөзқолданыс) өңдеу процесі жүйелі түрде жүргізіліп, нәтижесінде аз мезгіл аралығында (1963-1986 жж.) тек XIX-XX ғг. тілінің 90 млн. сөзқолданыстан тұратын мәтіні өңдеуден өтіп, 175 мың лексемадан тұратын сөзтізбе түзілген. Сонымен бірге, жаңадан құрастырылып жатқан француз тілінің 15 томдық сөздігіне (Treésor général de la Langue Francaise) 70 мың сөздік бірлігі мен 1 млн. иллюстрациялық цитаталар таңдалып алынған. Мұндағы әрбір сөздің қолдануы 10 жылдық уақыт мерзімімен құжатталған деуге болады. Аталған сөздіктің бірінші томының 1972 жылы жарық көруі, ал 15 томның толық аяқталуы 1987 жылға жоспарлануы сөздік құрастыру жұмысының автоматты және жүйелі түрде жүргізілгендігінің куәсі. Шынында да, бұдан бұрын ешбір ұлттық лексикография мұндай қарқынмен және сапамен сөздік құрастырмағаны белгілі.

Әрине, бұл жерде ҚазССР ҒА Тіл білімі институтында 1974-1986 жылдар аралығында дәстүрлі әдіспен (автоматсыз) құрастырылған көлемі 710 б/т болатын 10 томдық «Қазақ тілінің түсіндірме сөздігін» айта кеткен орынды. Аталған сөздікте барлығы 91 мың 502 лексикалық бірлік алынып, оларға 103 мың 2827 жазба нұсқадан жалпы саны 250 мыңға жуық мысал-үзінділер келтірілген. Ал қазіргі құрастырылып жатқан 15 томдық «Қазақ әдеби тілінің сөздігіне» жартылай автоматтандыру әдісі қолданылуда. Сөздікке 100 мыңнан аса сөздік бірлігі қамтылып, 5-6 жыл ішінде аяқталады деп жоспарланып отыр. Бұл сөздіктің негізгі мақсаты – «ұлттық әдеби тілдің сөз байлығын барынша жинақтау, әрбір сөздің мағыналарын, мағыналық реңктері мен стильдік ерекшеліктерін молынан қамту және олардың мазмұнын ашып көрсету; ұлттық әдеби тілдің қазіргі кезеңдегі нормалық сипатын таныту» [1, 7].

Сөздік ісінің қысқаша тарихын жалғастыра келе, Германияның Маннгейм қаласындағы неміс тілі институтында да аз көлемдегі сөздіктер құрастыру жұмысы қазіргі неміс тілінің жазба және ауызша дереккөздерінің мәшинелік картотекалары негізінде жүргізілгендігін айта кетуге болады. Мұндай лексикографиялық жұмыстар Швецияның Гетерборг университетінде, Мексика университетінде және т.б. елдерде де жүргізілуде.

Бұрынғы есептеуіш лексикография, ең алдымен, салалық лексикографияға ғана қатысты болатын. Яғни ол білім салаларын ақпаратпен

қамтамасыз ететін терминологиялық сөздіктердің компьютерлік пішінімен, сондай-ақ шағын лексиканы қамтитын автоматтанған аударма жүйелерімен әрекет ететін. Енді бұқаралық тұрғыда аударма мен редакциялық-баспа қызметін автоматтандыру мәшинелік сөздіктерге қойылатын талапты барынша күшейтті. Ол талап сөздіктердің саны мен сапасына тікелей қатысты қойылды.

Өткен ғасырдың 70 жылдарынан бастап бастама алған терминологиялық деректер банкі (ТДБ) және автоматты сөздіктер (АС) адам іс-әрекетін жеңілдетуге бағышталған болатын. Лексикографиялық жұмыстарды автоматтандыру құралдары барынша дамып, айтарлықтай дәрежеде есептеуіш лексикография арсеналын ұлғайта түсті және оны жалпы лексикография саласының айналысатын мәселелерімен барынша жақындатты.

Өткен ғасырдың 80-ші жылдары есептеуіш лексикографияның алдында жаңа ақпараттық технологияны меңгеруге байланысты тың мәселелер тұрған болатын. Ол табиғи тілде берілген кәсіби білім деректерін автоматты формалдауға қатысты өңдеу жұмыстарын жаңа технологиялық құралдарға – дербес компьютерлер мен ЭЕМ-нің 5-буынына және есептеуіш пен ақпараттық желілерді пайдалануға бағышталды. Аталған жайт жалпы лексикографияға бірнеше жаңа мәселелерді шешуді жүктеді. Нәтижесінде академиялық, салалық, есептеуіш және баспа лексикографияларын жекелемей, біртұтас қалпында қарастыру қажеттілігі артты. Себебі салалық лексикогафия табиғи тілдегі деректерді автоматты өңдеу жүйелерін пайдалануда модель түрінде көрініс табатын автоматтанған академиялық сөздіктерге мұқтаж еді. Баспа лексикографиясының алдында редекциялық-баспа процестерін автоматтандыру мен сол кездегі ақпарат тарату құралдарына сай келетін сөздік басылымдарының жаңа формасын іздеу мәселелері тұрғандықтан, ол аса дамыған есептеуіш лексикография құралдарын қажет етті.

1970 жылдардан бастап, лексикография қызметінің жақын онжылдықта жетекшілік ететін жаңа түрі айрықша анықтала бастаған. Ондай қызметке «адам-компьютер» арасындағы тілдік қатынасты лексикографиялық мәліметтермен қамтамасыз ету, оларды табиғи тілде өңдеу, ұлттық лексикографиялық қызмет көрсету жүйесін құру және т.б. жатқызамыз. Осыған байланысты «лингвистикалық құрылымдауға» қатысты есептеуіш лексикография болашағының рөлі мен орнын жаңаша ұғыну қажеттілігі туындады.

1986 жылы өзінің 20 жылдығын атап өткен есептеуіш лексикография, осыған дейін тек есептеуіш лингвистика бағыттарының бірі ғана болып саналған. Ал «есептеуіш лингвистика» термині 1966 жылдан бұрын да белгілі болғанымен, тек осы жылы ғана ол бірінші рет өзінің жеке зерттеу нысаны мен мәселелер ауқымы бар ғылыми пән деп аталды. Қазір оның құзыретіне автоматтанған лингвистикалық талдау (анализдеу), жинақтау (синтездеу) және мәтін мен сөйлеу тілін туындату, жасанды интеллект жүйесіндегі білім мәселесін шешу, тіл мен мәтін деректерін статистикалық деңгейде зерттеу,

тілдік зерттеулерді автоматтандыру мәселелерін қамтыды. Дәлірек айтсақ, сөздіктерді автоматты түрде дайындау, ақпараттық жүйелерді лингвистикалық мәліметтермен жабдықтау, мәшинелік аударма мен жасанды интеллектіні құрастыру мәселелері, редакциялық-баспа жүйелері мен ЭЕМ арқылы басқа да деректерді табиғи тілде өңдеуді көздейтін жағдаяттар.

Есептеуіш лингвистиканы арнайы филологтарға арналған пән ретінде оқытуға болатыны да белгілі. Яғни оны, есептеу техника пәні және оған бағдарлама жасау пәні деп қарастырып, жоғары оқу орындарындағы филолог мамандар үшін жеңілдетілген пішінде дәріс беруге қажеттілігін де айта кеткен орынды. Лингвистикалық құрылымдар мен есептеуіш құрылымдардың арақатынасын есептеуіш лингвистиканың зерттеу нысаны ретінде қарастырып, арнайы ғылыми пән ретінде қарастыруға болады.

Қазіргі кезде есептеуіш лексикография барлық тілдерге қатысты, соның ішінде қазақ тілі үшін де динамикалық сипатта дамып келе жатқан ғылыми бағыттардың бірі. Мұны тек теориялық және жүйелі түрдегі практикалық бағдарлама жасау жетістіктері ғана деуге болмайды, сондай-ақ оны тілдік процессорларды құрастыру аясындағы табыстармен де түсіндірген жөн. Бұл жердегі «процессор» сөзін компьютердің өзегі (миы) немесе негізгі бөлігі деп ұғынсақ, «тілдік процессор» сөз тіркесін – тілдік деректермен әрекет ететін компьютер деп, қарапайым түрде түсіндіреміз.

Жаңа технологиялық мүмкіндіктер аталған бағытты сөздіктердің автоматтанған картотекасы тұрғысынан жаңаша құрастыру қажеттігіне әкелді. Мәселен, сөздік картотекасына қатысты практикалық қажеттіктен туындаған мынадай әрекеттер орын алады: сөзтізбені дайындау, контекстерді таңдап алу мен іздеу, сөзқолданыстарды (бір немесе бір топ сөзге қатысты) салыстырмалы талдау, сөздік жобасына байланысты иллюстрациялық контекстер мен цитаталарды іріктеу, сөзге қатысты стильдік таңба жүйесін таңдау және т.б. [8].

Сонымен есептеуіш лексикография негізінде дамып келе жатқан қазақ тілінің сөздік құрастыру жұмысының әдістемесі қазіргі кездегі сөздік құрастыруда кездесетін көптеген кемшіліктерді қайталамауға бағытталған. Ондай кемшіліктерге жататындар: картотеканың бір жүйеге келтірілмеуі, сөздік мақаланы жазу ретінің әліпбилік ұстанымы, лексикалық инновацияларды тіркеудегі артта қалушылық, сөздіктердегі таңба-белгілердің ала-құлалығы (тіпті бір сөздік ішінде де) және т.б.

Есептеуіш лексикография технологиясы талдауға алынатын мәтіндегі сөзқолданыстардың тұтастығын сақтауда және сөздерді іріктеуде жүйелілік пен толықтықты қамтамасыз етеді. Басқару жүйесін деректердің мәтіндік базасына (ДМБ) қолдану мен оларға терминал дисплейі арқылы «ену» мүмкіндіктері сөздіктерді автоматтанған лексика-статистикалық анықтағыш ретінде әрекет етуіне жағдай жасайды. Автоматтанған лексика-статистикалық анықтағыш – ол әрі сөзнұсқағыш, әрі жиілік сөздік, әрі компьютер жадында тұрақты сақталатын мәтіндердің конкордансы.

Тұтынушы мәтіндік база мен бағдарламалық жабдықтауды өз қалауынша өзгертіп, толықтырып отыруына және қажетті деген материалды

немесе конкорданстың толық түрін (кері әліпбилі сөздікті де) қажетті реттілікпен қағаз бетіне басып шығара алады. Сонымен бірге ол тілдік деректерді редакциялау мен түзетулер жүргізудің жүйелік құралдарын пайдалану арқылы бастапқы, аралық және соңғы сатыдағы мәтіндерге кез келген түзетулер енгізуіне немесе оларды жаңадан туындатуына мүмкіндігі мол.

Қазіргі заманғы есептеуіш лексикография саласы сөздік жасайтын бағдарламалық құралдарды да құрастыра алады. Міне, мұны есептеуіш лексикографияның бірінші ерекшелігі дейміз. Ал екінші ерекшелігі – оның тілдік процессорларды құрылымдауға байланысты жетістіктері, яғни мәтіндерді автоматты талдау мен жинақтау бағдарламаларын қолдану мүмкіндігі. Есептеуіш лексикография тілдік процессорларды құрастыру ісіне әдеттегі сөздіктер мәліметтерін де нәтижелі қолданып келеді. Мәселен, орыс тілінің морфологиясы үшін мұндай жұмыс А.А.Зализняктың «Грамматический словарь русского языка» сөздігі бойынша орындалып, орыс тіліндегі сөзқолданыстарға морфологиялық анализ бен синтез жасайтын бірнеше компьютерлік бағдарламалар жазылды.

Есептеуіш лексикографияның үшінші ерекшелігі, оның әдеттегі сөздіктерді автоматтанған сөздікке айналдыру мен, керісінше, автоматты сөздіктерден кітап түріндегі сөздік алу мәселесін шешу мүмкіндігі. Автоматты сөздік (АС) деп, біз компьютер жадында сақталған әдеттегі сөздіктегі мақала пішінді ұстаным ететін сөздікті айтамыз. АС-ны әдеттегі сөздік ретінде пайдаланумен бірге, оны үздіксіз және қолма-қол түзетіп, толықтырып, жаңартып отыруға және оны келесі басылымға дайын қалпында сақтауға болады. Мәселен реестр сөздерге немесе оның фрагментіне (үзіндісіне) тіркелген белгілер бойынша бір топ сөздік мақаланы іздеуге болады; сөздіктің әртүрлі кескіндік бөліктерін жеке алып қарастыруға, яғни тек сөзтізбені (реестрді) ғана немесе сөздіктің түсіндірме бөлігін, немесе тек оның иллюстрациясын, немесе тек фразеологиялық бірліктерін, немесе тек сөздік мақалада орын алатын туынды сөздер тобын т.б.; сөздіктегі арнайы белгілер бойынша лексикалық топтарды іріктеп алып, әртүрлі көлемдегі сөздікшелерді құрастыруға да болады; әртүрлі сөздіктерді өзара салыстыра отыра, сөздік-дереккөздерінің макро-мен микроқұрылымдары бойынша бір-біріне қосып-алу әрекеттері арқылы сөздіктердің жаңа түрін құрастыруға мүмкіндік бар [8].

Қазіргі кезде компьютер оқый алатын қазақ тіліндегі мәтіндер саны да күннен-күнге көбейіп келеді. Бұл жайт (әлем бойынша алғанда) лексикография саласының нәтижелігін арттырып, қазақ тілінің автоматтанған картотекалық қорын құрастыруда, біртілді, екітілді және көптілді автоматты сөздіктерді жүзеге асыруға мүмкіндік тудырып отыр.

Сонымен қазақ тілінің лексикографиялық жұмыстарын есептеуіш техниканы қолданып автоматтандыру ісінде де есептеуіш лексикографияның жаңа даму кезеңі туды деуге болады.

Ә Д Е Б И Е Т

1. Қазақ әдеби тілінің сөздігі. Он бес томдық / Құраст. Т.Жанұзақов, С.Омарбеков, Ә.Жүнісбек және т.б. Алматы: Арыс, 2006. 1-том. 752 б.

2. *Busa R.* Index Thomisticus: Sancti Thomae Aquinatis Operum Omnium Indices et Concordantiae. Section Secunda. Concordantiae Operum Thomisticorum – Concordantia Prima. Stuttgart, 1974.

3. *Вертель В.А., Вертель Е.В., Рогожникова Р.П.* К вопросу об автоматизации лексикографических работ (некотрорые результаты применения ЭВМ) // ВЯ. 1978. № 2.

4. *Герд А.С., Богданов В.В., Азарова И.В.* и др. Автоматизация в лексикографии и словари-конкордансы // Филол. науки. 1981. №1.

5. *Джубанов А.Х.* Статистическое исследование казахского текста с применением ЭВМ (на материале романа М.Ауэзова «Абай жолы»: Автореф. дис. ... канд. Филол. Наук. Алма-Ата, 1973.

6. *Караулов Ю.Н., Молчанов В.И., Афанасьев В.А., Михалев Н.В.* Анализ метаязыка словаря с помощью ЭВМ. М.: Наука, 1982.

7. *Шайкевич А.Я.* Гипотезы о естественных классах и возможность количественной таксономии в современной лингвистике. М.: Наука. 1980.

8. *Андрющенко В.М.* Концепция и архитектура Машинного фонда русского языка. М.: Наука, 1989. 200 с.

ПРОФ. Қ. ЖҰБАНОВ ЖӘНЕ ҚОЛДАНБАЛЫ ТІЛ БІЛІМІНДЕГІ ФОРМАЛДЫ МОДЕЛЬДЕР

Құдайберген Жұбанов тағылымы: ғылыми сабақтастық және жаңа бағыттар: Республикалық ғылыми-практикалық конференция материалдары. – Алматы, 2009 ж. – 70-77 бб.

Профессор Қ.Жұбанов қазақ тілінің зерттелу болашағын өзіне тән көрегендікпен, табиғи дарындылықпен болжам жасап, олардың кейбір тұстары тілдің қолданбалы саласының жаңа бағыттарынан да байқалады. Дәлірек айтсақ, тіл зерттеу аясындағы ХХ ғ. 60-жылдарында ғана пайда болған «Статистикалық лингвистика», «Лингвистикадағы модельдеу» бағыттарының кейбір бастамаларын Құдайберген Жұбановтың сол кездегі зерттеулерінің өзінен-ақ молынан кездестіруге болады.

Қ.Жұбанов өткен ғасырдың 30-жылдарының өзінде-ақ өз зерттеулерін заманға сай жүргізуге ынталанғанын «Қазақ тілінің жоғарғы курсы» атты жазған еңбегіндегі мына тұжырымнан байқауға болады [1, 79-б.]: «XIX ғасырда жаратылыс ғылымы ілгеріледі; оларды зерттеуге түрлі метод қолданылды. Объектив әдісімен іс істеу тілшілерге де жұқты. Бұрын тіл логика негізді әдіспен зерттелетін еді (яғни, сағызша созыла беретін белгілі өлшеуге келмей, көбіне субъектив жорамалға сүйенетін)», – дейді.

Ғалымның сол кездің өзінде-ақ тіл білімінде сандық деректерді пайдаланудың қажеттілігіне де ерекше ден қойғаны байқалады. Мәселен,

ғалымның 1936 жылы жарық көрген «Қазақ тілінің грамматикасы» атты еңбегінде сөз мүшелерінің орны, дыбыстардың саны мен сапасы, түрлері туралы жазған тұжырымы осыған дәлел: «Дыбыстың саны бар да, сапасы бар. Дыбыстың сапасы бірдей бола тұрып, саны әлденешеу бола береді. … Бір тілде жұмсалатын дыбыстардың жалпы санының ұшы-қиыры жоқ, бірақ оның бәрінің бірдей сапасы әртүрлі бола бермейді. Сапасы әртүрлі болатын дыбыстар белгілі мөлшерде ғана болады да, көп болмайды. … Қазақ тілінде де сапасы әртүрлі болатын дыбыстардың арнаулы саны бар…» [1, 183-б.]. Тілдік құбылыстарда сан мен сапаны ажырату қажеттігін айта келе, ғалым тұрпат (форма) пен мағынаны да ажырата білудің маңыздылығын тиянақтап айта білген. Мысалы, «Қазақ тілінің жоғарғы курсында» былай дейді [1, 79-б.]: «…Бірақ (грамматикалық құрылымдардың) сыртқы белгілеріне сүйенуден формализм туды. Сөйтіп морфология жеке сөздердің формасын, синтаксис сөз тіркестерінің формасын тексеретін ғылым деп танылды. … Синтаксистің өзі де … тіркескен сөздердің тіркесу формасын қарайтын ғылым болды» – дей келе, бұл әдісті толық қолдамайтынын білдіреді. Осындай күмәнін дәлелдей түсу мақсатымен және ол бағытты толығынан терістемейтінін білдіру үшін ғалым «Қазақ тілі грамматикасының кейбір мәселелері» атты еңбегінде тағы да былай дейді [1, 34-б.]: «…Рас, біз формашыларға қарсымыз: тіл құбылыстарын жалаң түр жағын ғана зерттеумен біліп болмайды. Бұл – сөзсіз сыңаржақтық. Сонда да жоғынан бары игі. Түр – форма деген тілде жоқ белгі емес, бар нәрсе. Бар қасиетін көрсетушілік қашанда пайда болмаса, зиян келтірмейді. Бірақ жеткіліксіз, өйткені түр (форма) тілдің бар тарауы емес, бір тарауы».

Шындығында да, ғалым зерттеулерінде тілдің формалық сипатына және объективтілікке сүйенетін жақтары баршылық.

Енді, қазіргі кезде қолданбалы тіл білімінің дәстүрлі бағытымен қатар дамып келе жатқан және заман сұранысына сай пайда болған оның жаңаша бағытындағы «тілді модельдеу» әдісінің жоғарыдағы тақырыпшаларда сөз болған кейбір теориялық тұстарын қайталай отыра, салыстырмалы пайымдаулар жасауға әрекет етейік.

Модельдеу әдісінің қажеттігі, ең алдымен, зерттеу нысанымызды тікелей байқауға, яғни қолмен ұстап, көзбен көруге болмайтын жағдайлардан туындайды. Сондықтан осындай модельдеу әдісі, ол өзінше бір бөлек әдіс ретінде, көптеген ғылымдар саласында, соның ішінде қазіргі кезде қазақ тілін зерттеуде де қолданыс тауып жүргені белгілі.

Ғылыми әдебиеттерде «модель» терминіне мынандай анықтама беріледі: «Модель дегеніміз – зерттеу нысанының қасиеттерін бойына сақтай алатын және түпнұсқаның (нысанның) орнына жүретін, танымдық процесінің қажеттігінен туындайтын материалдық не абстракты нысан (объект)» делінген.

Модель әр уақытта өзінің түпнұсқа түріндегі шындығынан да артық күйде, яғни «дәріптелген» сипатта қарастырылады. Осының негізінде, қай ғылым саласының болсын, формалды модельдері математикалық жүйе деп қарастырылады. Сондықтан, басқаша айтқанда, формалдылықтың түсінігі –

математикалық дәлдік немесе біріздiлік деуге болады. Бұл қасиеттер тіл табиғатына да тән құбылыс.

Осымен қатар, лингвистикалық модель эксплонаторлық қасиетке де ие болу керек деп ұйғарылады, яғни оның түсініктеме беру мүмкіндігі болуы қажет. Аталған қасиетке ие болуы үшін, ол, ең алдымен, осыған дейінгі тілдік теория негізінде түсінігі күңгірт не түсініксіз болып жүрген тілдік құбылыстарды бірізді түсіндіре алатын, анықтама бере алатын мүмкіндігі болуы керек. Сол сияқты, зерттеу нысанының бұрын-соңды ескерілмей жүрген жаңа мүмкіндіктеріне де ғылыми болжам жасай алатын қабілеті болуы қажет деп пайымдалады.

«Лингвистикалық модельдеуде» сөйлеу қызметі моделінің пішінін екі түрде қарастырады. Оның біріншісі – семантикасыз немесе тек грамматиканы игеруге ғана «еліктеуші» таза синтаксистік модель. Екіншісі – мағыналы сөйлемдерді түсіндіре білумен қатар, оларды дұрыс құра да білу мүмкіндігіне еліктейтін – семантикалық модельдер.

Мұндай модельдерді құру жазба не сөйлеу мәтіндерін автоматты түрде өңдеуде аса маңызды. Мәселен, «машиналық аударма» жүйелерінің барлық дамыған түрлерінде «автоматтандырылған синтаксистік талдау» жүргізу сөйлемнің семантикалық түсінігін алу үшін және екінші тілдік жүйедегі сөйлемдер құрылымын «аударма грамматикасына» ауыстыру үшін аса қажет деп саналады. Сол сияқты, «ақпараттық іздеу» жүйесінде де толық синтаксистік талдау сөйлем құрамындағы «есімді сөз тіркестері» негізінде танылады. Ал «жасанды интеллект» жасау мәселесіндегі мәтінді тану жүйесіне қатысты синтаксистік талдау, өзінше бір бөлек семантика-синтаксистік топтаманы құрайды деуге болады.

Қазіргі кезде, қай тілдің болмасын, оған тән сөйлем құрылымының дәл және формалды берілісін білу аса қажет болуда. Ол үшін реттелген сызықтық тізбек құрайтын кесінділер жиыны – **«құрастырушылар жүйесін»** немесе **«құрастырушылар моделін»** құра білу керек. Аталған модельде синтаксистік бірлік ретінде сөзтұлғалар, сөз тіркестер және олардың құрылымдық қатынастары қарастырылады [2].

Жоғарыдағы тақырыпшаларда баяндалғандай, жүйе жиынындағы қатынастар «граф» деп аталатын сызба түрінде көрініс табады да және оның төбешігіндегі ең кіші бірлік ретінде не негіз сөз, не сөзтұлға, не сөз тіркес алынады. Сөйлемдердің графтық сызбасын салудың алдында сөйлемдегі әрбір сөзқолданыстар өз реті бойынша нөмірленеді. Әрі қарай, сызбалар түрінде граф төбешігіндегі құрастырушы сөзқолданыстар мағыналық жағынан ішкі сөз тіркестерге бөлектеліп, олар бір-бірімен ішкі және сыртқы тәуелділіктеріне қарай, түзу сызықтық кесінділер арқылы өзара қосылады. Осындай сызба пішіні әдебиеттерде **«құрастырушылар моделі»** немесе **«синтаксистік модель»** деген терминдермен аталып жүр [2].

Сөйлем құрастырушыларына талдау жүргізудің өзіндік қиыншылығына қарамай, **«құрастырушылар модельдері»** сөйлем құрылымын грамматикалық және семантикалық тұрғыда да дұрыс сипаттайды деп саналады. Мұндай дұрыстық сипат, көбінде, мағына ажырату кезінде,

мысалы, құрастырушы сөзтұлғалар – **«омонимдес сөздер»** немесе сөйлемдер – **«омонимдес сөйлемдер»** болған жағдайларда анығырақ байқалатыны дәлелденеді [2].

Сөйлемдер ішіндегі сөзтұлғалар арасындағы **қабысу, жанасу, қиысу, меңгеру** сияқты тәуелділіктер жиыны сол сөйлемнің синтаксистік құрылымы жайлы көптеген мәліметтер бере алады. Ал тәжірибе жүзіндегі ізденістерде, мұндай тәуелділіктердің жиыны граф пішініндегі сызба түрінде көрініс табады.

Профессор Құдайберген Жұбановтың сөйлем құрылымына қатысты зерттеулерінде сөйлем ішіндегі сөзтұлғалар мен сөз тіркестер «граф» немесе «құрастырушылар моделі» деген терминдермен аталмаса да, сөйлемге енетін сөздердің әрбіреуін тиісті нөмірлермен таңбалау және олардың тәуелділіктерін сызба түрінде кесінділер мен төртбұрыштар арқылы беру молынан кездеседі.

Мәселен, ғалымның «О формах сочетания слов в казахском языке» атты мақаласында [1, 476-б.] 15 сөзқолданыстан тұратын мындай сөйлемді қарастырады: *«Сен(1) көрген(2) қымбат(3) алпауыттан(4) да(5) Қызыл(6) құрт(7) Ертай(8) ұстаның(9) Темірбайдан(10) алған(11) төрт(12) бесті(13) аты(14) жақсы(15)».* Осы сөйлемдегі сөздердің орын ауысуынан сөйлем мағынасының өзгеру заңдылығын көрсету мақсатымен, әрбір сөзді нөмірлеп алып, бес түрлі айтылым ретінде бөлек-бөлек қарастырып, пайымдаулар жасайды. Сөйлемді орыс тіліне былайша аударады: *«Чем(4) дорогой(3) помещик(4) виденный(2) тобою(1), лучше(15) и(5) гнедой(12) пятилетний(13) мерин(14), полученный(11) кузнецом(9) Ертаем(8) кызыл-куртовцем(6) от(7) Темирбая(10)».*

Әрі қарай осы сөйлемдегі сөздерді түпнұсқадағы баламаларымен салыстыра отыра, олардың мағыналары мен тұлға өзгерістерін талдайды. Міне, мұндай әрекеттер қазіргі кезде арнайы әдіс ретінде сөйлем синтаксисін модельдеуге қатысты зерттеулерде жиі қолданыс тауып жүр.

Қ.Жұбановтың зерттеулерінде, яғни **«Сөйлем мүшелері»** [1, 153-157-бб.], **«Жетек сөз бен жетекші сөз»** [1, 160-161-бб.], **«Сөйлем мүшелерінің орын-орны»** [1, 162-б.] деген тақырыпшаларда **«О формах сочетания слов в казахском языке»** [1, 476-481-бб.], **«О построении речи на казахском языке»** [1, 482-495-бб.] және **«Этюды к переводу трех стилей»** [1, 531-534-бб.] атты мақалаларында көрініс тапқан сызбалар үлгісі 12 түрлі сурет арқылы берілген. Бұл келтірілген сызбаларды қолданбалы лингвистика терминімен айтсақ, белгілі мағынада, **«Қазақ сөйлемдерінің синтаксистік модельдерінің үлгілері»** деуге де болады.

Енді сөйлем ішіндегі тәуелділіктердің кесінді сызба түрінде көрініс табуын Қ.Жұбановтың аталған кітабындағы мынадай пайымдауларынан байқауға болады:

1) кітаптың 156-бетінде *«Айқындауышты өзінің айқындап тұрған сөзімен сызық арқылы байланыстырып көрсетсек мынадай болады»* деп екі түрлі сызбаны ұсынады;

2) кітаптың 160-бетінде – *«Біздер мектепке келдік»* деген сөйлемді мысал ретінде алып, былай пайымдаған: «… бұл үш сөз матастырылмаған. *Біздер* деген мен *мектепке* деген екеуі жиналып *келдік* деген бір сөзге арқандалған. Мұның арқандалу ретін сызықпен көрсетсек мынадай болады» деп және *«Енді қай сөз, қай сөзді жетектеп тұрғанын да сызбамен таңбаласақ, мынадай болады»* деп сызба түрінде ұсынады;

3) кітаптың 162-бетіндегі «Сөйлем мүшелерінің орын-орны» деген тақырыпшада бастауыш, баяндауыш, анықтауыш және толықтауыш болатын сөйлем мүшелерінің орындарын көрсету үшін М.Әуезов шығармасынан алынған мына сөйлемнің *«Бұл – мал күткен, малға ықылас берген бейілдің көрмесі»* әр сөзін нөмірлеп алып, былай дейді: «Мұндағы сөздердің қалайша қиындасып тұрғанын сызықпен көрсетсек, мынандай болады» деп сызбаға сілтеме жасайды.

4) кітаптың 482 бетіндегі «О построении речи на казахском языке» атты мақаласында 16 сөзқолданыстан тұратын сөйлемді қарастырып, олардың формалды белгілері арқылы және бір сөзге қатысты параллель тіркестердің орын тәртібінің қатал сақталмай, тек логикалық екпінге ғана бағыныштылығы жайлы сызбаны ұсынады. Бұл сызбаның сөйлемнің синтаксистік моделінің көрінісі екендігіне осы саланың мамандары күмән туғызбас деген ойдамыз.

5) келесі қарастыратынымыз ғалым кітабының 531-бетіндегі көлемі жағынан шағын (4 бет), ал құндылығы жағынан «аударманың синтаксистік және семантикалық моделі» ғылыми бағыты үшін аса маңызды «Этюды к переводу трех стилей» атты (тақырыбы шартты – *ред.*) атты зерттеу жұмысы. Бұл мақала ҚазССР-ның Орталық архивінен алынып, тұңғыш рет Алматыда «Лексико-семантические вопросы перевода» атты жинақта 1987 жылы басылып шыққан болатын. Мұның құндылығы үш түрлі стиль мәтіндеріндегі (көркем-әдебиет, ғылыми және публицистикалық) сөйлемдерді салыстыра-салғастыра қарастыруында. Әр стильге тән ерекшеліктерге қатысты сөйлемдер ішіндегі сөздердің бір-бірімен байланысын 5 түрлі сызбалар арқылы орыс тіліндегі сөйлемдерді қазақ тіліне аударуға қатысты өзіндік семантика-синтаксистік модель-үлгіні ұсынуында.

Ғалым Құдайберген Жұбанов аталған сызбаларына сілтеме жасау үшін мынандай айтылымдарды қолданған: *«сызық арқылы байланыстырып көрсетсек», «мұның арқандалу ретін сызықпен көрсетсек», «қай сөз, қай сөзді жетектеп тұрғанын да сызбамен таңбаласақ», «сөздердің қалайша қиындасып тұрғанын сызықпен көрсетсек»* және т.б.

Тілді оның формалды пішіні арқылы сипаттау «алмасу ақпаратының» мазмұндық сырын ашуға толық мүмкіндік туғыза бермейтіндіктен зерттеуге тиісті ақпараттың мағыналық, яғни семантикалық жағын бөлек қарастыру қажеттігі туындайды.

Жоғарыдағы тақырыпшаларда сөз болғандай, көптеген зерттеушілердің тұжырымдауынша, сөйлем мағынасы аса күрделі және көпаспектілі құрылым. Сондықтан синтаксистік модельдеу барысында ұстаным етілетін төрт түрлі ғылыми бағыттың үшеуі «әлем–ойлау–тіл» үшбұрышы

элементтерінің қатыстылығына қарай, ал төртіншісі – «тіл–адам» прагматикалық қатынасына негізделеді. Ғылыми әдебиеттерде бұл бағыттар: *онтологоцентристік, концептоцентристік, синтактикоцентристік* және *антропоцентристік бағыт* деген терминдермен аталып жүр.

Міне, осындай көптеген терминдер арқылы аталып жүрген «қолданбалы тіл білімінің» жаңа бағытына қатысты мәселелер Қ.Жұбановтың зерттеулерінде дәл олай аталмағанымен, өзіндік мазмұнда молынан көрініс тапқан.

Жоғарыда қысқаша баяндалған мәселелер қазақ тілінің қолданбалы саласының жаңа бағыты бойынша жүргізілетін зерттеулерде өз жалғасын табатыны мәлім.

Әдебиет

1. *Жұбанов Қ. Қ.* Қазақ тілі жөніндегі зерттеулер. – Алматы: «Ғылым», 1999, – 581 б.

2. *Жұбанов А.Қ.* Қолданбалы лингвистика: формалды модельдер: Оқу құралы. – 2 бас. – Алматы: Қазақ университеті, 2007. – 280 б.

ПРОФ. Қ. ЖҰБАНОВ ЗЕРТТТЕУЛЕРІ – ҚОЛДАНБАЛЫ ТІЛ БІЛІМІНІҢ ЖАҢА САЛАСЫНА ҒЫЛЫМИ БАҒЫТ-БАҒДАР

«Жұбанов тағылымы VII» халықаралық ғылыми конференция материалдары (профессор Қ.Қ.Жұбановтың 110 жылдық және белгілі лингвист ғалым, профессор Е.Қ.Жұбановтың 80 жылдық мерейтойларына арналған). I том. –Ақтөбе, Қ.Жұбанов атындағы АқМУ РББ, 2009. –28-31-бб.

Заман сұранысына сай жоғары дәрежелі жүйелерді зерттеуде ғалымдар өздерінің ғылыми көзқарастарына сапалық өзгерістер енгізіп, теорияның логикалық құрылысын жаңарту, яғни адам және қоғам туралы ғылым салаларына ойлау мен дүниетану жолдарына тың өзгерістер енгізулері керек болды. Себебі ойлау стилі – біріншіден, инженерлік-математикалық, ал екіншіден, ықтималдық сипатта болуы шарт екені анықталды [1].

Классикалық ғылымның табиғи-ғылымдық көрінісінің болашақ заңды дамуының жалғасы ретіндегі инженерлік-ықтималды стильдің ғылыми ойлаудағы айырмашылығы – оның сырттан қабылданатын ақпаратты «кванттарға» бөлшектеуінде, яғни ақпараттың ұсақ бөліктерін математикалық логика әдістерінің көмегімен біртұтас жүйе түрінде қайтадан туындату мүмкіншілігінде; көпшілік жағдайда оның абсолюттеу идеясынан бас тартуында; дискретті математика әдістері көмегімен әртүрлі табиғи процестерді модельдеуінде; модельденген процестерді техникалық құралдар арқылы (компьютер көмегімен) жүзеге асыруында болып отыр. Осы

айтылғандардың бәрі дерлік белгілі бір жағдайда казіргі лингвистикалық ойлау стиліне қатысы бар.

Құдайберген Қуанұлы Жұбанов өткен ғасырдың 30-жылдарының өзінде-ақ өз зерттеулерін заманға сай жүргізуге ұмтылған ғалым. Мәселен, оны ғалымның «Қазақ тілінің жоғарғы курсы» атты еңбегіндегі мынадай тұжырымынан байқауға болады: «XIX ғасырда жаратылыс ғылымы ілгеріледі; оларды зерттеуге түрлі метод қолданылды. Объектив әдісімен іс істеу тілшілерге де жұқты. Бұрын тіл логика негізді әдіспен зерттелетін еді (яғни, сағызша созыла беретін белгілі өлшеуге келмей, көбіне субъектив жорамалға сүйенетін)», – дейді [2, 79-б.].

Ғалымның осындай зерттеу тұстары тілдің қолданбалы саласының жаңа бағыттарынан да байқалады. Дәлірек айтсақ, тіл зерттеу аясындағы XX ғ. 60-жылдарында ғана пайда болған «Статистикалық лингвистика», «Лингвистикадағы модельдеу» бағыттарының кейбір бастамаларын Құдайберген Жұбановтың сол кездегі зерттеулерінен кездестіруге болады. Мысалы, қазіргі кездегі тіл білімінде, соның ішінде қазақ тіл білімінде де «Лингвистикалық статистика» деп аталып жүрген жаңа саланың алғашқы нышандары мен белгілерінің ғалым еңбектерінде көрініс тапқаны белгілі. Осыған байланысты математик ғалым, филология ғылымдарының докторы, профессор Қ.Б.Бектаев былай деген болатын: «...сандық деректер мен мәліметтерді, оны пайдалану нәтижесіндегі кейбір пікірлер мен тұжырымдарды біз проф. Қ.Жұбановтың еңбектерінен жиі кездестіреміз. Ғалымның сол кездің өзінде-ақ тіл білімінде сандық деректерді пайдалану кажет екендігіне ерекше ден қойғанын айқын байкаймыз» [3].

Ғалым Қ.Жұбановтың 1936 жылы жарық көрген «Қазақ тілінің грамматикасы» атты еңбегінде сөз мүшелерінің орны, дыбыстардың саны мен сапасы, түрлері туралы жазған тұжырымы да осыған дәлел: «Дыбыстың саны бар да, сапасы бар. Дыбыстың сапасы бірдей бола тұрып, саны әлденешеу бола береді. … Бір тілде жұмсалатын дыбыстардың жалпы санының ұшы-қиыры жоқ, бірақ оның бәрінің бірдей сапасы әр түрлі бола бермейді. Сапасы әр түрлі болатын дыбыстар белгілі мөлшерде ғана болады да, көп болмайды. … Қазақ тілінде де сапасы әр түрлі болатын дыбыстардың арнаулы саны бар…» [2, 183-б.].

Шын мәніндегі статистикалық ізденістерді Қ.Жұбанов қазақ орфографиясын ғылыми негізде құру мақсатында қолданған. Ол әліпби құрамына енуге тиісті әріптердің санын қазақ тіліндегі белгілі бір фонеманы іс жүзіндегі қажеттігіне, оған деген ділгерлікке, қолданыс жиілігіне бағындыра белгілеуді мақсат тұтқан [2, 114-б.].

Ғалым тілдің сандық сипаты негізінде тілдік бірліктердің сапалық қасиеттерін де анықтауға болатынын сөз ете отырып, ол сандық деректердің негізінде тілге қатысты кейбір жайттар (дыбыстар, буындар, сөздер, сөйлем мүшелерінің орны және т.б.) бойынша тың тұжырымдар жасаған. Бұлардың маңыздылығы сонда – олар қазіргі кезде қолданбалы тіл білімінің жаңа саласындағы ізденуші-зерттеушілерге үлгі-тірек, бағыт-бағдар болып отыр [3, 115-116-бб.].

Тілдік құбылыстардағы сан мен сапа жайында айта келе, ғалым тұрпат (форма) пен мағынаны да ажырата білудің маңыздылығын түсіне білген. Мысалы, «Қазақ тілінің жоғарғы курсы» атты еңбегінде ғалым былай дейді [2, 79-б.]: «…Бірақ (грамматикалық құрылымдардың) сыртқы белгілеріне сүйенуден формализм туды. Сөйтіп морфология жеке сөздердің формасын, синтаксис сөз тіркестерінің формасын тексеретін ғылым деп танылды. … Синтаксистің өзі де … тіркескен сөздердің тіркесу формасын қарайтын ғылым болды», – дей келе, бұл әдісті толық қолдай бермейтінін де білдіреді. Осы мәселеге күмән туғызған тұстарын дәлелдей түсу мақсатымен және ол бағытты толығынан терістей де қоймайтынын білдіру үшін ғалым «Қазақ тілі грамматикасының кейбір мәселелері» атты еңбегінде тағы да мынадай ой-тұжырымдарын айтады: «…Рас, біз формашыларға қарсымыз: тіл құбылыстарын жалаң түр жағын ғана зерттеумен біліп болмайды. Бұл – сөзсіз сыңаржақтық. Сонда да жоғынан бары игі. Түр – форма деген тілде жоқ белгі емес, бар нәрсе. Бар қасиетін көрсетушілік қашанда пайда болмаса, зиян келтірмейді. Бірақ жеткіліксіз, өйткені түр (форма) тілдің бар тарауы емес, бір тарауы» [2, 34-б.]. Шындығында да, ғалым зерттеулерінде тілдің формалық сипатына және объективтілігіне сүйенетін жақтары баршылық.

Профессор Құдайберген Жұбановтың түсіндіруінше, тіліміздегі қарым-қатынасқа түскен сөздердің мағыналық жағын танып қана қою жеткіліксіз, сол мағына қалайша туындап тұр, қандай формалардың әсері болды, яғни мағына мен форманың аралық тығыз қарым-қатынасын ескеріп отыру қажеттігіне оқырманның назарын аудартады. «Жаңа грамматиканың жаңалықтары жайында» атты еңбегінде ғалым былай дейді: «Грамматиканың мақсаты жалғыз-ақ сөздің қандай мағыналары болатындығын баяндау ғана емес, қандай формаларға қандай мағыналар сәйкес келетінін баяндап, дұрысы, қандай мағына шығу үшін сөзді қалайша құрастыру керек екендігін баяндау. Сондықтан құрастыру, құрылыс жағын елемейтін грамматика да болмақ емес» [2]. Ғалым «құрастыру», «құрылыс» (структура) ұғымдары арқылы грамматиканың үлкен бір саласы – синтаксисті осылайша түсіндіреді.

«Өткен ғасырдың 50-жылдарынан бастап грамматикада өзінше бір жеке сала болып бөлініп шыккан сөз тіркесі синтаксисі де Қ.Жұбановтың назарынан тыс қалмады» [4], – деп жазады синтаксис саласының белгілі маманы проф. Қ.Есенов. Аталған маңызды мәселеге арнап проф. Қ.Жұбанов «О формах сочетания слов в казахском языке» атты мақала жазғаны бізге мәлім. Ғалым еңбектеріндегі тағы бір көңіл бөлерлік жайт – сөйлем мүшелерінің орналасу тәртібіндегі заңдылықтар және кісі аттары мен біріккен сөздер табиғатының кеңінен сөз болуы. Қысқаша айтқанда, профессор Қ.Жұбанов қазақ тілінің өзге салаларымен бірге тіліміздің синтаксистік құрылысын зерттеуде үлкен теориялық еңбек қалдырып, жаңа бағыттағы болашақ зерттеулерге аса білгерлікпен жол салып кетті.

Қазіргі қоғам элементтерінің арасындағы өте күрделі байланыстарды анықтау қоғам өміріндегі ақпарат пен ақпараттау жүйесін басқарудың рөлін арттырады. Осындай күрделі жүйеге мәтінді де жатқызуға болады. Қай

тілдегі мәтін болмасын, онда: құрылымдық-коммуникативтік және номинативтік, әлеуметтік және индивидуальдік, статикалық және динамикалық, детерминдік және ықтималдық т.б. да әртүрлі аспектілердің араласып жататыны белгілі. Тек осы аталған факторларды ескерсек қана ойымызда туындайтын мәтін мазмұнынан өзара байланыстағы сөйлем тізбектеріне (мәтінге) өтудің заңдылықтарын ашуға мүмкіндік жасалады. Бұл жағдайды зерттеп-тану, мәтін құрайтын бірліктерді бөлшектеу арқылы талдау (анализдеу) немесе мәтінді құрастыруға қажетті бірліктерді жинақтау (синтездеу) әдісі арқылы ғана жүзеге асады демекпіз.

Синтездеу әдісін қолдану кезінде әртүрлі сөздерді тізбектеп, сөйлем құрастыру, сол сөйлемдердің тізбектері негізінде мәтін абзацтарын құру және т.б. мәтіндердің ірі бөлшектерін (параграф, тарау т.б.) құрастырудың заңдылықтары ашылады. Осы заңдылықтарды танып-білу, бір тілден екінші тілге компьютер арқылы аударуды, автоматты түрде мәтіннен реферат алу мен аннотациялауды және қажетті деген деректерді автоматты түрде тез арада іздеп-табу проблемадарына басқаша көзқараспен қараумен бірге, белгілі үлгімен жасалған құжат қағаздарын синтездеу мәселесіне де көп жеңілдік жасалатыны сөзсіз. Тек осындай зерттеулердің нәгізінде ғана тілдің әртүрлі деңгейдегі қызметінде байқалатын ережелер мен ережеден тыс жағдайлардың «шындыққа тән немесе тән емес» екендігін тексеруге мүмкіндік туады.

Компьютер арқылы алынған ғылыми нәтижелер алғашында қазақ тіліндегі мәтіндердің мазмұнын автоматты түрде ашуға көмектессе, болашақта мамандардың компьютермен табиғи тілде, яғни қазақ тілінде де қатынас жасап, «дидарласуына» мүмкіндіктер туғызары сөзсіз.

Енді қазіргі кезде қолданбалы тіл білімінің дәстүрлі бағытымен қатар дамып келе жатқан және заман сұранысына сай пайда болған оның жаңаша бағытындағы «тілді модельдеу» әдісінің қажеттігі, ең алдымен, зерттеу нысанымызды тікелей байқауға, яғни оны «қолмен ұстап, көзбен көруге болмайтын» жағдайлардан туындайды. Сондықтан осындай модельдеу әдісі, өзінше бір бөлек әдіс ретінде, көптеген ғылымдар саласында, соның ішінде қазіргі кезде қазақ тілін зерттеуде де қолданыс тауып жүргені белгілі.

Ғылыми әдебиеттерде «модель» терминіне келесідей анықтама берілген: «Модель дегеніміз – зерттеу нысанының қасиеттерін бойына сақтай алатын және түпнұсқаның (нысанның) орнына жүретін, танымдық процестің қажеттігінен туындайтын материалдық не абстракты нысан (объект)» делінген. Модель әр уақытта өзінің түпнұсқа түріндегі шындығынан да артық күйде, яғни «дәріптелген» сипатта қарастырылады. Осының негізінде, қай ғылым саласының болсын, формалды модельдері математикалық жүйе деп қарастырылады. Сондықтан, басқаша айтқанда, формалдылықтың түсінігі – математикалық дәлдік немесе біріздiлік деуге болады. Бұл қасиеттер тіл табиғатына да тән құбылыс деуге болады.

Сөйлемдер ішіндегі сөзтұлғалар арасындағы **қабысу, жанасу, қиысу, меңгеру** сияқты тәуелділіктер (байланыстар) жиыны сол сөйлемнің синтаксистік құрылымы жайлы көптеген мәліметтер бере алады. Ал тәжірибе

жүзіндегі ізденістерде, мұндай тәуелділіктер (байланыстар) жиыны граф пішініндегі сызба түрінде көрініс табады.

Профессор Құдайберген Жұбановтың сөйлем құрылымына қатысты зерттеулерінде сөйлем ішіндегі сөзтұлғалар мен сөз тіркестер «граф» немесе «құрастырушылар моделі» деген терминдермен аталмаса да, сөйлемге енетін сөздердің әрбіреуін тиісті сандық белгілермен таңбалау және олардың тәуелділіктерін сызба түрінде кесінділер мен төртбұрыштар арқылы беру молынан кездеседі. Мәселен, ғалымның «О формах сочетания слов в казахском языке» атты мақаласында 15 сөзқолданыстан тұратын сөйлемдегі сөздердің орын ауысуынан сөйлем мағынасының өзгеру заңдылығын көрсету мақсатымен, әрбір сөзді нөмірлеп алып (тиісті сандармен таңбалап), бес түрлі айтылым ретінде бөлек-бөлек қарастырып, пайымдаулар жасайды [2, 476-б.]. Әрі қарай осы сөйлемдегі сөздерді түпнұсқадағы баламаларымен салыстыра отыра, олардың мағыналары мен тұлға өзгерістерін талдайды. Міне, мұндай әрекеттер қазіргі кезде арнайы әдіс ретінде сөйлем синтаксисін модельдеуге қатысты зерттеулерде жиі қолданыс тауып жүр.

Қ.Жұбановтың «Сөйлем мүшелері» [2, 153-157-бб.], «Жетек сөз бен жетекші сөз» [2, 160-161-бб.], «Сөйлем мүшелерінің орын-орны» [2, 162-б.] деген тақырыпшаларда, «О формах сочетания слов в казахском языке» [2, 476-481-бб.], «О построении речи на казахском языке» [2, 482-495-бб.] және «Этюды к переводу трех стилей» [2, 531-534-бб.] атты мақалаларына қатысты зерттеулерінде көрініс тапқан сызбалар үлгісі 12 түрлі сызба арқылы берілген. Бұл келтірілген сызбаларды қолданбалы лингвистика терминімен айтсақ, белгілі мағынада, **«Қазақ сөйлемдерінің синтаксистік модельдері»** деуге де болар еді.

Ғалым Құдайберген Жұбановтың аталған зерттеулеріндегі сызбаларда қажетті сілтеме жасау мақсатында мынандай айтылымдар қолданыс тапқан: *«сызық арқылы байланыстырып көрсетсек», «мұның арқандалу ретін сызықпен көрсетсек», «қай сөз, қай сөзді жетектеп тұрғанын да сызбамен таңбаласақ», «сөздердің қалайша қиындасып тұрғанын сызықпен көрсетсек»* және т.б.

Көптеген зерттеушілердің тұжырымдауынша, сөйлем мағынасы аса күрделі және көпаспектілі құрылым. Сондықтан синтаксистік модельдеу барысында ұстаным етілетін төрт түрлі ғылыми бағыттың үшеуі «әлем – ойлау – тіл» үшбұрышы элементтерінің қатыстылығына қарай, ал төртіншісі – «тіл – адам» прагматикалық қатынасына негізделеді. Ғылыми әдебиеттерде бұл бағыттар: *онтологоцентристік, концептоцентристік, синтактико-центристік* және *антропоцентристік бағыт* деген терминдермен аталып жүр. Міне, осындай толып жатқан терминдер арқылы аталып жүрген «Қолданбалы тіл білімінің» жаңа бағытына қатысты мәселелер Қ.Жұбановтың зерттеулерінде (келтірілген терминдермен аталмаса да) өзіндік мазмұнда молынан көрініс тапқан.

Жоғарыда қысқаша баяндалған мәселелер қазақ тілінің қолданбалы саласының жаңа бағыты бойынша жүргізілетін зерттеулерде өз жалғасын табады деп сенім білдіруге болады.

Сонымен, қазақ тілінің теориялық және қолданбалы тіл білімінің негізін қалаушы, қазақ филологиясының тұңғыш профессоры Құдайберген Қуанұлы Жұбановтың азғана өмірінің бір-ақ мақсаты болған сияқты. Ол – әлі ашылмаған шындықты ашу немесе соның ашылуына ықпал жасау, сол арқылы қоғамға қызмет ету. Мұндай мақсатқа жетудің әртүрлі жолдары бар екені белгілі, бірақ ғалым Қ.Жұбанов солардың ішінен зерттеудің дәлдігі мен объективтілігін қамтамасыз ететін тәсілді ғана ең дұрысы деп тапқан болу керек. Шындығында да, дәлдік пен объективтілікке ұмтылу – қазіргі ғылымдарға бірден-бір тән қасиет, ал бұның қазақ тіл біліміне де қатысы бары сөзсіз.

Әдебиет

1. *Пиотровский Р.Г.* Инженерная лингвистика и теория языка. – Л., 1979. – 112 с.
2. *Жұбанов Қ.* Қазақ тілі жөніндегі зерттеулер. – Алматы: Ғылым, 1999. –581 б.
3. *Бектаев Қ., Белботаев А., Смаилов Т.* Қазақ тіл біліндегі лингвистикалық статистиканың алғашқы көріністері // Қ.Жұбанов және қазақ совет тіл білімі. – Алматы, 1990. – 108-116-бб.
4. *Есенов Қ.* Профессор Қ.Жұбанов және қазақ тілінің синтаксисі // Құдайберген Жұбанов және қазақ совет тіл білімі. – Алматы, 1990. – 71-б.
5. *Жубанов Х.* Исследование по казахскому языку. – Алма-Ата, I966. – С. 220.

ПРОФ. Қ.ЖҰБАНОВ ЖӘНЕ ҚОЛДАНБАЛЫ ТІЛ БІЛІМІНІҢ ЖАҢА БАҒЫТЫ

ҚР Білім және ғылым министрлігі Ғылым комитеті
А.Байтұрсынұлы атындағы Тіл білімі институты
«Қазақ тіл білімінің өзекті мәселелері» атты
профессор Құдайберген Жұбановтың 110 жылдығына
арналған халықаралық ғылыми-теориялық
конференция материалдары. Алматы-2009. 27-32-бб.

Профессор Қ.Жұбановтың филологияға қосқан теориялық және практикалық ғылыми үлестері қазақ тіл білімінің барлық салаларына әсер етті деуге болады. Қазақ тілінің зерттелу болашағына өзіне тән көрегендікпен, табиғи дарындылықпен болжам жасап, олардың кейбір тұстарының ғалым зерттеулерінде орын алғанына арнайы тоқталған жөн. Ғалымның осындай зерттеу тұстары тілдің қолданбалы саласының жаңа бағыттарынан да байқалады. Дәлірек айтсақ, тіл зерттеу аясындағы XX ғ. 60-жылдарында ғана пайда болған «Статистикалық лингвистика», «Лингвистикадағы модельдеу» бағыттарының кейбір бастамаларын

Құдайберген Жұбановтың сол кездегі зерттеулерінің өзінен-ақ молынан кездестіруге болады.

Қ.Жұбанов өткен ғасырдың 30-жылдарының өзінде-ақ өз зерттеулерін заманға сай жүргізуге ұмтылғанын «Қазақ тілінің жоғарғы курсы» атты еңбегіндегі мына тұжырымнан байқауға болады: «XIX ғасырда жаратылыс ғылымы ілгеріледі; оларды зерттеуге түрлі метод қолданылды. Объектив әдісімен іс істеу тілшілерге де жұқты. Бұрын тіл логика негізді әдіспен зерттелетін еді (яғни, сағызша созыла беретін белгілі өлшеуге келмей, көбіне субъектив жорамалға сүйенетін)», – дейді [1, 79-б.].

Ғалым сол кездің өзінде-ақ тіл білімінде сандық деректерді пайдаланудың қажеттілігіне ерекше көңіл бөледі. Сол сияқты, тіл білімінің белгілі ғалымдары қатарында Қ.Жұбанов та тілді зерттеуде, тілдік фактілерді пайымдауда сандық деректерге, ұғымдарға жүгінгендігін ғалым зерттеулерінен байқауға болады. Қазақ тіл біліміндегі «Лингвистикалық статистика» деп аталып жүрген жаңа саланың алғашқы нышандары мен белгілерінің ғалым еңбектерінде көрініс тапқандары туралы белгілі ғалым, филология ғылымдарының докторы, математик Қ.Б.Бектаев былай дейді: «...сандық деректер мен мәліметтерді, оны пайдалану нәтижесіндегі кейбір пікірлер мен тұжырымдарды біз проф. Қ.Жұбановтың еңбектерінен жиі кездестіреміз. Ғалымның сол кездің өзінде-ақ тіл білімінде сандық деректерді пайдалану кажет екендігіне ерекше ден қойғанын айқын байкаймыз» [2]. Мәселен, ғалымның 1936 жылы жарық көрген «Қазақ тілінің грамматикасы» атты еңбегінде сөз мүшелерінің орны, дыбыстардың саны мен сапасы, түрлері туралы жазған тұжырымы осыған дәлел: «Дыбыстың саны бар да, сапасы бар. Дыбыстың сапасы бірдей бола тұрып, саны әлденешеу бола береді. … Бір тілде жұмсалатын дыбыстардың жалпы санының ұшы-қиыры жоқ, бірақ оның бәрінің бірдей сапасы әр түрлі бола бермейді. Сапасы әр түрлі болатын дыбыстар белгілі мөлшерде ғана болады да, көп болмайды. … Қазақ тілінде де сапасы әр түрлі болатын дыбыстардың арнаулы саны бар…». Сол сияқты Ілияс Жансүгіровтың шығармалары бойынша алған статистикалық деректерін былайша пайымдайды: «… Ілиястың небәрі 12 жол, «Қойшы ойы» деген кішкене өлеңінде 488 дыбыс бар. Бірақ, бұл 488 дыбыс 488 түрлі емес, мұнда не бары 28-ақ түрлі дыбыс бар. Осы 28 түрлі дыбыстың бірі 10, бірі 50 рет қайталап барып 488 болған. … Қазақ тілінде де сапасы әр түрлі болатын дыбыстардың арнаулы саны бар» [1, 183-б.].

Профессор Қ.Жұбановтың «Қазақ тілі жөніндегі зерттеулерінде» бұл жайлардан басқа да мәселелер, айталық: сөз тіркесінің формалары, сөйлем мүшелерінің орын тәртібі, буындар және ондағы дыбыс саны, буындарды оқыту әдістемесі т.б. мәселелер сөз болған тұста көптеген сандық (статистикалық) деректер мен мәліметтер келтірілген [1].

Шын мәніндегі статистикалық ізденістерді Қ.Жұбанов қазақ орфографиясын ғылыми негізде құру мақсатында қолданған. Ол әліпби құрамына енуге тиісті әріптердің санын қазақ тіліндегі белгілі бір фонеманы іс жүзіндегі қажеттігіне, оған деген ділгерлікке, қолданыс жиілігіне бағындыра белгілеуді мақсат тұтқан [2, 114-б.].

Міне, бұл айтылғандар – проф. Қ.Қ.Жұбановтың тілді зерттеуде, оның заңдылықтарын қарауда есеп-санаққа, ретті жерінде математикалық ойлауға, оның тәсілдеріне сүйенгендігінің тағы да бір дәлелі.

Қ.Жұбанов еңбектеріндегі сандық деректерді пайдалану жайымен таныса отырып, мынадай қорытынды жасауға болады: проф. Құдайберген Қуанұлы Жұбановтың қазақ тіл білімі туралы зерттеулерінде тілді сандық деректер мен мәліметтер негізінде қараудың алғашқы нышандары бар. Бүгінде тіл білімінің бір саласына айналып отырған «Лингвистикалық статистиканы» ғалым сол кездің өзінде-ақ тани білген. Тілдің сандық сипатынан байқалатын сапалық қасиеттерін оны зерттеуде пайдалануға болатын тиімді әдіс көздерінің бірі екенін де байқаған. Сөйтіп, ол сандық деректердің негізінде тілге қатысты бірсыпыра жайларды (дыбыстар, буындар, сөздер, сөйлем мүшелерінің орны т.б.) нақты анықтап, кейбір тың тұжырымдар жасаған. Бұлардың маңыздылығы сонда – олар қазіргі кезде бұл саладағы ізденуші-зерттеушілерге үлгі-тірек, бағыт-бағдар болып отыр [2, 115-116-бб.].

Тілдік құбылыстарда сан мен сапаны ажырату қажеттігін айта келе, ғалым тұрпат (форма) пен мағынаны да ажырата білудің маңыздылығын тиянақтап айта білген. Мысалы, «Қазақ тілінің жоғарғы курсында» ғалым былай дейді [1, 79-б.]: «…Бірақ (грамматикалық құрылымдардың) сыртқы белгілеріне сүйенуден формализм туды. Сөйтіп морфология жеке сөздердің формасын, синтаксис сөз тіркестерінің формасын тексеретін ғылым деп танылды. … Синтаксистің өзі де … тіркескен сөздердің тіркесу формасын қарайтын ғылым болды», - дей келе, бұл әдісті толық қолдай қоймайтынын білдіреді. Осындай күмән туғызған тұстарын дәлелдей түсу мақсатымен және ол бағытты толығынан терістей де қоймайтынын білдіру үшін ғалым «Қазақ тілі грамматикасының кейбір мәселелері» атты еңбегінде тағы да былай дейді: «…Рас, біз формашыларға қарсымыз: тіл құбылыстарын жалаң түр жағын ғана зерттеумен біліп болмайды. Бұл – сөзсіз сыңаржақтық. Сонда да жоғынан бары игі. Түр – форма деген тілде жоқ белгі емес, бар нәрсе. Бар қасиетін көрсетушілік қашанда пайда болмаса, зиян келтірмейді. Бірақ жеткіліксіз, өйткені түр (форма) тілдің бар тарауы емес, бір тарауы» [1, 34-б.]. Шындығында да, ғалым зерттеулерінде тілдің формалық сипатына және объективтілікке сүйенетін жақтары баршылық.

Профессор Құдайберген Жұбановтың түсіндіруінше, тіліміздегі қарым-қатынасқа түскен сөздердің мағыналық жағын танып қана қою жеткіліксіз, сол мағына қалайша туындап тұр, қандай формалардың әсері болды, яғни мағына мен форманың аралық тығыз қарым-қатынасын ескеріп отыру қажеттігіне оқырманның назарын аудартады. «Жаңа грамматиканың жаңалықтары жайында» атты еңбегінде ғалым былай дейді: «Грамматиканың мақсаты жалғыз-ақ сөздің қандай мағыналары болатындығын баяндау ғана емес, қандай формаларға қандай мағыналар сәйкес келетінін баяндап, дұрысы, қандай мағына шығу үшін сөзді қалайша құрастыру керек екендігін баяндау. Сондықтан құрастыру, құрылыс жағын елемейтін грамматика да болмақ емес» [1]. Ғалым «құрастыру», «құрылыс» (структура) ұғымдары

арқылы грамматиканың үлкен бір саласы синтаксисті осылайша түсіндіреді, өткен ғасырдың 50-жылдарынан бастап грамматикада өзінше бір жеке сала болып бөлініп шыккан сөз тіркесі синтаксисі де Қ.Жұбановтың назарынан тыс қалмады [4], - деп жазады синтаксис саласының белгілі маманы проф. Қ.Есенов. Осы маңызды мәселеге арнап проф. Қ.Жұбанов «О формах сочетания слов в казахском языке» атты мақала жазғаны да бізге мәлім. Тағы бір көңіл бөлерлік жайт – ғалымның еңбектерінде сөйлем мүшелерінің орналасу тәртібіндегі заңдылықтар, сондай-ақ кісі аттары мен біріккен сөздердің табиғаты кеңінен сөз болуы. Қысқаша айтқанда, профессор Қ.Жұбанов қазақ тілінің өзге салаларымен бірге тіліміздің синтаксистік құрылысын зерттеуде үлкен теориялық еңбек қалдырып, жаңа бағыттағы болашақ зерттеулерге аса білгерлікпен жол салып кетті.

Қолданбалы тіл білімінің «Мәтін лингвистикасы», «Мәтін теориясы» атты жаңа бағытқа қатысты да Қ.Жұбановтың жоғарыда айтылған және т.б. зерттеулерінің маңызы зор. Қазіргі кезде ғалымдар мәтін құрайтын элементтердің заңдылықтарын ашу жолдарын қарастырып, оны талдау (анализ) әдісінен гөрі, жинақтау (синтездеу) әдісімен көбірек шұғылданатынын байқауға болады. Оның басты себебі, компьютерлік лингвистика әдістерімен мәтіннің «туындау» заңдылықтарын анықтау, мәтін мазмұнын автоматты түрде ашу, мәтіндерді бір тілден екінші тілге аудару, мәтіннен аннотация мен реферат алу қажеттігі және т.б. қолданбалы мәселелердің шешімін табу болып отыр.

Мәтіндік деңгейдегі зерттеулер – түркі тіл білімінде (оның ішінде қазақ тіл білімінде) әлі де болса өз деңгейінде қолға алынбай жүрген мәселенің бірі. Мәтін грамматикасының кейбір теориялық мәселелері Н.З.Гаджиеваның, К.М.Абдуллаевтың жұмыстарында кездеседі. Түркі тідерінің синтаксистік құрылымын зерттеуде аналитизмге бас бұру тенденциясы алғашында проф. Қ.Жұбановтың, бертін келе Н.С.Поспеловтың, Б.А.Серебренниковтің, Н.О.Оралбаеваның және Н.Х.Демесинованың және т.б. ғылыми ізденістерінен кездестіруге болады.

Мәтіндердің туындау заңдылықтарын анықтау, әсіресе олардың ішінен мағына жағынан ең негізгі деген тілдік бірліктерді ажырата білудің, қазақ тілін үйрету (үйрену) үрдісінде де тәжірибелік мәні зор. Ол үшін ең әуелі мәтіндерді сегменттеу ережелерінің көмегімен үлкенді-кішілі мәтіндік бірліктерге топтау әдісін жүргізудің қажеттігі туады. Зерттеушілердің тәжірибесінде мәтіннің құрылымына қатысты бірліктер ретінде: сөздер, сөз тіркестері, сөйлемдер, абзацтар, параграфтар, бөлімдер, тараулар, томдар алынып жүр.

Ғалым Қ.Жұбановтың қазақ тіліндегі сөз тіркестеріне аса көңіл бөлуі сөз тіркестерінің синтаксис саласының пайда болуына үлкен себін тигізгені белгілі. Оның сөз тіркесіне байланысты пікірінің құндылығы, сөздердің өзара байланысуында жарыспалы (параллельное) және тізбекті (последовательное) амалдарын ашып керсетуінде. Сонымен бірге сөйлем мүшелерінің бірінші, екінші, үшінші және т.б. қатарда кездесетіндігі жайлы пікірлерінің де мәні зор. Сөйлем ішіндегі сөздердің біразы бір-бірімен тікелей тіркессе, енді

біреулерінің баспалдақтап тіркесу мүмкіндігін де атап өтеді. Осындай құнды деректердің біздің ізденістерімізге де қажеттігі сөзсіз.

Енді тілші-ғалымдардың кейінгі зерттеулерінің нәтижесі ретінде (қазақ тілінің сөз тіркесі синтаксисінің пайда болуы тәріздес) қазақ тілінің жаңа саласы – мәтін синтаксисіне әкелуі де күмән туғыза қоймас.

Синтаксистік зерттеулер сөйлемнің барынша күрделенген түрлерімен ғана шектелмей, зерттеу шеңберін кеңейтіп, мәтін құрылымына бағытталып, оның негізгі деген бірліктерінің және олардың арасындағы күрделі байланыстардың сыр-сипаты мен шарттарын белгілеуі өзінше бір проблемаға айналып, мәтін мазмұнын оның тақырыбына тікелей байланыстыра ашу қажеттігі туды. Осы мәселені шешу барысында мәтінге тән жалпы тақырыпты іштей шағын тақырыпшаларға (подтема, субподтема, микротема) ажырату қолайлы деп есептелінді.

Автордың белгілі бір жайды оқырманға тұжырымдап айтып жеткізуі кезінде сөздерді орын-орнымен, рет-ретімен таңдай білудің жолдарын анықтау да өзінше бір күрделі мәселе. Сол себепті біздер өз зерттеулерімізде сөздер жиынтығын үлкен екі топқа бөліп, олардың біріншісін – мағыналық сөздер тобы, ал екіншісін – мәнерлеуші сөздер тобына жатқызып қарастырғанды жөн көрдік. Мағыналық топтағы сөздердің сөйлемдегі синтаксистік құрылысты қалайша толықтыра түсу жағын өз алдына зерттей келе, оның абзац пен мәтін көлеміндегі сөздердің жиілігі мен сөздердің өзара байланыс «күшіне» де қатысты екенін, ол жағдайдың арнайы жазылған логикалық бағдарламалар аркылы іске асатындығын да атап айтқымыз келеді.

Табиғат пен қоғамда кездесетін оқиғалар мен құбылыстарды құрайтын жоғары дәрежелі жүйелердің заңдылықтарын зерттеуде ғалымдарға енді өз іс-әрекеттеріне жаңаша тұрғыда қараудың қажеттігі туындап отыр. Мұндағы жоғары дәрежелі жүйелер деп отырғанымыз – тірі организмдер, қоғам және табиғи тіл. Бұларды зерттеу, олардың құрылымдық жағынан көп деңгейлігін, детерминдігі мен кездейсоқтықтың сіңісуін, өзара бөлшектерінің белгілі-бір дербестігі мен үйлесімді тұтастығын ескеретін жаңа теорияны жете зерттеу бүгінгі таңның күн тәртібінің бірден-бір мәселесі болып отыр. Мұндай жоғары дәрежелі жүйелерді зерттеуде ғалымдар өздерінің ғылыми көзқарастарына сапалық өзгерістер енгізіп, теорияның логикалық құрылысын жаңарту қажеттігі туындады. Басқаша айтқанда, адам және қоғам туралы ғылыммен айналысатын ғылым салаларына ойлау мен дүниетану жолдарына тың өзгерістер енгізілуі керек болды. Олай дейтініміз, ойлау стилі – біріншіден, **инженерлік-математикалық**, екіншіден, онда **ықтималдықтың** да болуы шарт екені анықталды [5]. Біз оны – **инженерлік-ықтималдық** стиль деп атауды жөн көрдік.

Классикалық ғылымның табиғи-ғылымдық көрінісінің болашақ заңды дамуының жалғасы ретіндегі инженерлік-ықтималды стильдің ғылыми ойлаудағы айырмашылығы – оның сырттан қабылданатын ақпаратты «кванттарға» бөлшектеуінде, яғни ақпараттың ұсақ бөлшектерін математикалық логика әдістерінің көмегімен біртұтас жүйе түрінде қайтадан

туындату мүмкіншілігінде; көпшілік жағдайда оның абсолюттеу идеясынан бас тартуында; дискретті математика әдістері көмегімен әртүрлі табиғи процестерді модельдеуінде; модельденген процестерді техникалық құралдар арқылы (компьютер көмегімен) жүзеге асыруда болып отыр.

Осы айтылғандардың бәрі дерлік белгілі бір жағдайда казіргі лингвистикалық ойлау стиліне қатысы бар.

Тіл және оның атқаратын қызметі тұрғысынан алғанда, өмір талабынан туындап отырған, жоғарыда аталған лингвистика саласының алдындағы осы мақсат, тілге деген жүйелілік-құрылымдық көзқараспен қараудың шешімі толық табыла бермейді. Қазіргі қоғам элементтерінін арасындағы өте күрделі байланыстарды анықтау қоғам өміріндегі ақпарат пен ақпараттау жүйесін басқарудың рөлін арттырады. Осындай күрделі жүйеге мәтінді де жатқызуға болады. Қай тілдегі мәтін болмасын, онда: құрылымдық-коммуникативтік және номинативтік, әлеуметтік және индивидуальдік, статикалық және динамикалық, детерминдік және ықтималдық т.б. да әртүрлі аспектілердің араласып жататыны белгілі. Тек осы аталған факторларды ескерсек қана ойымызда «пайда болатын» мәтін мазмұнынан өзара байланыстағы сөйлем тізбектеріне (яғни мәтінге) өтудің заңдылықтарын ашуға мүмкіндік туады. Бұл жағдайды зерттеп-тану, мәтін құрайтын бірліктерді бөлшектеу арқылы талдау (анализдеу) немесе мәтінді құрастыруға қажетті бірліктерді жинақтау (синтездеу) әдісі арқылы ғана жүзеге асады демекпіз.

Екінші жағдайда, яғни синтездеу әдісін қолдану кезінде әртүрлі сөздерді тізбектеп, сөйлем құрастыру, сол сөйлемдердің тізбектері негізінде мәтін абзацтарын құру және т.б. мәтіндердің ірі бөлшектерін (параграф, тарау т.б.) құрастырудың заңдылықтары ашылады. Осы заңдылықтарды танып-білу, бір тілден екінші тілге компьютер арқылы аударуды, автоматты түрде мәтіннен реферат алу мен аннотациялауды және қажетті деген деректерді автоматты түрде тез арада «іздеп-табу» проблемадарына басқаша көзқараспен қарау үшін ғана емес, сонымен бірге, тіпті белгілі үлгімен жасалған құжат қағаздарын синтездеу мәселесіне де көп жеңілдік келтіретіні сөзсіз. Тек осындай зерттеулердің негізінде ғана тілдің әртүрлі деңгейдегі қызметінде байқалатын ережелер мен ережеден тыс жағдайлардың «шындыққа тән немесе тән емес» екендігін тексеруге мүмкіндік туады.

Компьютер арқылы алынған ғылыми нәтижелер алғашында қазақ тіліндегі мәтіндердің мазмұнын автоматты түрде ашуға көмектессе, болашақта мамандардың компьютермен табиғи тілде, яғни қазақ тілінде де қатынас жасап, «дидарласуына» мүмкіндіктер туғызары сөзсіз.

Енді қазіргі кезде қолданбалы тіл білімінің дәстүрлі бағытымен қатар дамып келе жатқан және заман сұранысына сай пайда болған оның жаңаша бағытындағы «тілді модельдеу» әдісінің қажеттігі, ең алдымен, зерттеу нысанымызды тікелей байқауға, яғни оны қолмен ұстап, көзбен көруге болмайтын жағдайлардан туындайды. Сондықтан осындай модельдеу әдісі, ол өзінше бір бөлек әдіс ретінде, көптеген ғылымдар саласында, соның ішінде қазіргі кезде қазақ тілін зерттеуде де қолданыс тауып жүргені белгілі.

Ғылыми әдебиеттерде «модель» терминіне мынандай анықтама беріледі: «Модель дегеніміз – зерттеу нысанының қасиеттерін бойына сақтай алатын және түпнұсқаның (нысанның) орнына жүретін, танымдық процестің қажеттігінен туындайтын материалдық не абстракты нысан (объект)» делінген. Модель әр уақытта өзінің түпнұсқа түріндегі шындығынан да артық күйде, яғни «дәріптелген» сипатта қарастырылады. Осының негізінде, қай ғылым саласының болсын, формалды модельдері математикалық жүйе деп қарастырылады. Сондықтан, басқаша айтқанда, формалдылықтың түсінігі – математикалық дәлдік немесе бірізділік деуге болады. Бұл қасиеттер тіл табиғатына да тән құбылыс деуге болады.

Сөйлемдер ішіндегі сөзтұлғалар арасындағы **қабысу, жанасу, қиысу, меңгеру** сияқты байланыстар жиыны сол сөйлемнің синтаксистік құрылымы жайлы көптеген мәліметтер бере алады. Ал тәжірибе жүзіндегі ізденістерде, мұндай байланыстардың жиыны граф пішініндегі сызба түрінде көрініс табады.

Профессор Құдайберген Жұбановтың сөйлем құрылымына қатысты зерттеулерінде сөйлем ішіндегі сөзтұлғалар мен сөз тіркестер «граф» немесе «құрастырушылар моделі» деген терминдермен аталмаса да, сөйлемге енетін сөздердің әрбіреуін тиісті сандармен таңбалау және олардың тәуелділіктерін сызба түрінде кесінділер мен төртбұрыштар арқылы беру молынан кездеседі. Мәселен, ғалымның «О формах сочетания слов в казахском языке» атты мақаласында [1, 476-б.] 15 сөзқолданыстан тұратын сөйлемдегі сөздердің орын ауысуынан сөйлем мағынасының өзгеру заңдылығын көрсету мақсатымен, әрбір сөзді нөмірлеп алып (тиісті сандармен таңбалап), бес түрлі айтылым ретінде бөлек-бөлек қарастырып, пайымдаулар жасайды. Әрі қарай осы сөйлемдегі сөздерді түпнұсқадағы баламаларымен салыстыра отыра, олардың мағыналары мен тұлға өзгерістерін талдайды. Міне, мұндай әрекеттер қазіргі кезде арнайы әдіс ретінде сөйлем синтаксисін модельдеуге қатысты зерттеулерде жиі қолданыс тауып жүр.

Қ.Жұбановтың зерттеулерінде, яғни «Сөйлем мүшелері» [1, 153-157-бб.], «Жетек сөз бен жетекші сөз» [1, 160-161-бб.], «Сөйлем мүшелерінің орын-орны» [1, 162-б.] деген тақырыпшаларда, «О формах сочетания слов в казахском языке» [1, 476-481-бб.], «О построении речи на казахском языке» [1, 482-495-бб.] және «Этюды к переводу трех стилей» [1, 531-534-бб.] атты мақалаларында көрініс тапқан сызбалар үлгісі 12 түрлі сызба арқылы берілген. Бұл келтірілген сызбаларды қолданбалы лингвистика терминімен айтсақ, белгілі мағынада, **«Қазақ сөйлемдерінің синтаксистік модельдері»** деуге де болады.

Ғалым Құдайберген Жұбанов зерттеулеріндегі келтірген сызбаларына сілтеме жасау үшін мынандай айтылымдар қолданыс тапқан: *«сызық арқылы байланыстырып көрсетсек», «мұның арқандалу ретін сызықпен көрсетсек», «қай сөз, қай сөзді жетектеп тұрғанын да сызбамен таңбаласақ», «сөздердің қалайша қиындасып тұрғанын сызықпен көрсетсек»* және т.б.

Көптеген зерттеушілердің тұжырымдауынша, сөйлем мағынасы аса күрделі және көпаспектілі құрылым. Сондықтан синтаксистік модельдеу

барысында ұстаным етілетін төрт түрлі ғылыми бағыттың үшеуі «әлем–ойлау–тіл» үшбұрышы элементтерінің қатыстылығына қарай, ал төртіншісі – «тіл–адам» прагматикалық қатынасына негізделеді. Ғылыми әдебиеттерде бұл бағыттар: *онтологоцентристік, концептоцентристік, синтактико-центристік* және *антропоцентристік бағыт* деген терминдермен аталып жүр. Міне, осындай толып жатқан терминдер арқылы аталып жүрген «қолданбалы тіл білімінің» жаңа бағытына қатысты мәселелер Қ.Жұбановтың зерттеулерінде дәл солай аталмағанымен, өзіндік мазмұнда молынан көрініс тапқан.

Жоғарыда қысқаша баяндалған мәселелер қазақ тілінің қолданбалы саласының жаңа бағыты бойынша жүргізілетін зерттеулерде өз жалғасын табады деп сенім білдіруге болады.

Сонымен, қазақ тілінің теориялық және қолданбалы тіл білімінің негізін қалаушы, қазақ филологиясының тұңғыш профессоры Құдайберген Қуанұлы Жұбановтың азғана өмірінің бір-ақ мақсаты болған сияқты. Ол – әлі ашылмаған шындықты ашу немесе соның ашылуына ықпал жасау, сол арқылы қоғамға қызмет ету. Мұндай мақсатқа жетудің әртүрлі жолдары бар екені белгілі, бірақ ғалым Қ.Жұбанов солардың ішінен зерттеудің дәлдігі мен объективтілігін қамтамасыз ететін тәсілді ғана ең дұрысы деп тапқан болу керек. Шындығында да, дәлдік пен объективтілікке ұмтылу – қазіргі ғылымдарға бірден-бір тән қасиет, ал бұның қазақ тіл біліміне де қатысы бары сөзсіз.

Әдебиет

1. *Жұбанов Қ.* Қазақ тілі жөніндегі зерттеулер. – Алматы: Ғылым, 1999. –581 б.

2. *Бектаев Қ., Белботаев А., Смаилов Т.* Қазақ тіл біліміндегі лингвистикалық статистиканың алғашқы көріністері // Қ.Жұбанов және қазақ совет тіл білімі. – Алматы, 1990. – 108-116-бб.

3. *Жубанов Х.* Исследование по казахскому языку. – Алма-Ата, I966. – С. 220.

4. *Есенов Қ.* Профессор Қ.Жұбанов және қазақ тілінің синтаксисі // Құдайберген Жұбанов және қазақ совет тіл білімі. – Алматы, 1990. – 71-б.

5. *Пиотровский Р.Г.* Инженерная лингвистика и теория языка. – Л., 1979. – 112 с.

IV БӨЛІМ

ҚАЗАҚ ТІЛІНІҢ ҰЛТТЫҚ МӘТІНДЕР КОРПУСЫН ҚҰРУДЫҢ АЛҒЫШАРТТАРЫ

***

ПРЕДПОСЫЛКИ СОЗДАНИЯ НАЦИОНАЛЬНОГО КОРПУСА ТЕКСТОВ КАЗАХСКОГО ЯЗЫКА

ҚАЗАҚ ТІЛІ МӘТІНДЕР КОРПУСЫНЫҢ КОМПЬЮТЕРЛІК БАЗАСЫН ҚҰРУДЫҢ АЛҒЫШАРТТАРЫ

ҚР БжҒМ ҒК А.Байтұрсынұлы атындағы Тіл білімі институты
«Әдеби тіл және қазақ тілінің өміршеңдігі» атты ҰҒА академигі,
филология ғылымдарының докторы Рәбиға Сыздықтың 85
жылдығына орай өткізілген ғылыми-теориялық конференция.
Алматы-2009. 173-179-бб.

Соңғы жылдары «Корпустық лингвистика» саласы ғылыми айналымға айтарлықтай дәрежеде енді деуге болады. Тілдік зерттеулерде ды тәжірбие жүзінде пайдалануда, әсіресе сөздіктер мен неше түрлі грамматикаларды дайындау кезінде, мәтіндік корпустардың аса құнды материал болатындығы анықталып отыр. Қазіргі кезде корпустық лингвистиканың мәселелері кейбір оқу құралдарының да арнайы тақырыбына айналуда [1; 2].

Қазіргі қазақ тіл білімінің даму үдерісі заман өзгерісіне сай, жаңа технологияны – компьютерлер мүмкіндіктерін ұлттық тілді зерттеу барысында кең түрде қолдануды қажет етуде. Осыған байланысты А.Байтұрсынұлы атындағы Тіл білімі институтында «Компьютерлік лингвистика» мен «Корпустық лингвистика» салаларына тікелей қатысты бар «Мәдени құндылықтар ретіндегі қазақ тіліндегі мәтіндер корпусы және сөздіктердің «Тіл – қазына» атты ұлттық компьютерлік қоры» атты тақырыпқа қатысты зерттеу жұмысы бастама алып отыр. Ал бұл зерттеу жұмысының негізгі мақсаты – қазақ тілінің мәдени құндылығы болып саналатын толық мәтіндеріне қажеттікке сай грамматикалық белгі-кодтар енгізіп, оның дербес түрдегі «Тіл – қазына» атты мәтіндер корпустарының

компьютерлік базасын құру. Аталған база коммуникация қажеттігін толығымен қанағаттандыра алатындай, үлкен көлемді мәтіндер мен сөздіктерге қатысты ақпараттық-сұраныс жүйесі түрінде көрініс табуы керек. Алғашында (2009-2011 ж.ж.) толық мәтіндердің компьютерлік қорының нысандары ретінде М.Әуезовтің, Ә.Кекілбаевтың, М.Мақатаевтың, М.Мағауинның толық шығармаларынан тек таңдама мәтіндер ғана алынады деп жоспарлануда.

Қазақ тілінің лексикалық қорының үнемі өсіп отыру үдерісін лексикографиялық ізденістер мен баспадан шығып жатқан біртілді, екітілді сөздіктер түрлерінен де байқауға болады. Сол себепті, қазақ тілін зерттеуде корпустық лингвистика саласына көп мән беріп, оның әлемдік дәрежедегі теориялық және практикалық жақтарын күн тәртібіне қойып, арнайы зерттеу жұмыстарын жүргізу қажет демекпіз.

Әлем бойынша мәтін корпустарын құрастыру мен олардың қызметіне қатысты жалпы және нақты мәселелерге арналған мақалалар жарық көрген ғылымы журналдардың арнайы басылымдары да шыға бастағаны мәлім [3; 4].

Осы саланың мамандары корпустық лингвистиканы тілдік корпустарды құру мен оны пайдалану жағдайын зерттейтін тіл білімінің бір бөлімі ретінде қарастырып келеді. Сонымен бірге, ол таңдама мәтіндер (мәтін үзінділері) түрінде көрініс табатын тілді зерттеу мен оны танып білуге қажетті маңызды материал деуге де болады. Кейбір ғалымдар ол пәннің түсінігін тар шеңберде қарастырып, оны тек компьютерлік лингвистика саласының аясында ғана түсіндіреді [5]. Ал компьютерлік лингвистика ұғымын, әдетте, компьютерлік құралдарды пайдаланудың кең мүмкіндігі ретінде түсіндіруге болатыны белгілі. Бұл жердегі «компьютерлік құралдар» деп отырғанымыз – компьютерлік бағдарламалар, тілдік деректерді өңдеу мен компьютерлік технологияны орынды ұйымдастыру жұмыстары және т.б. [6, 13-б.]. Міне, сондықтан да корпустық лингвистика өзіне жүктелген міндетті ондай құралсыз атқара да алмас еді. Бірақ компьютер мұндай рөлді қазіргі білім саласының барлық түрлерінде де атқаратынын ескерсек, онда олардың бәрін бірдей компьютерлік лингвистика саласына жатқыза беруге болмайды. Сондықтан біз корпустық лингвистиканың қазіргі тілтаным парадигмасында алатын орнын анықтау үшін оның тарихына қысқаша тоқталайық.

Корпустық лингвистика Н.Хомскийдің көзқарасына қарама-қарсы бағытта пайда болды деуге әбден болады. Себебі, оның оппозициялық «құзырет/түсінігі» («компетенция/представление») бойынша тіл жайындағы деректер тіл иесінің құзыретін құрайды да, сондықтан да әртүрлі аса көлемді тілдік нысанды зерттеудің орнына сөйлеушінің тілдік интуициясына (ішкі түйсігіне) сүйенгенін жеткілікті деп санайды. Корпустық лингвистика мен оған қатысты пікірталастар, айтарлықтай дәрежеде, теоретикалық және эмпирикалық білім аяларының ең ұтымды арақатынастарын іздестіру жолын бейнелейді деуге болады.

Корпустық лингвистика 1963 ж. АҚШ-та өмірге келіп, Браун корпусынан (The Brown Standard Corpus of American English) бастама алады.

Бұл корпустың көлемі 1 млн. сөзқолданыстан тұрады және оның құрамында әрбіреуі 2 мың сөзқолданысқа тең 500 мәтін бар [7]. Браундық корпус осыған ұқсас корпустар құруға қатысты зерттеулердің кеңінен тараған (атышулы) нысаны мен стандартына айналды. Ғалымдар арасында көптеген лингвистикалық зерттеулерді сапалы жүргізу тек ауқымды тілдік материалдар негізінде ғана жүзеге асатындығын ұғыну байқала бастады. Осы айтылғандардың барлығы мәтіндерді корпус түрінде ұйымдастыру ережелері мен оларға талдау жүргізу әдіснамасын зерттейтін бағыттың пайда болуына себепші болып, ғалымдарды одан әрі ынталандыра түсті. Сонымен корпустық лингвистика осы әдіс-тәсілдердің әдіснамасы ретінде туындап отыр деуге әбден болады.

Корпустық лингвистиканың ағылшын тілі материалында туындауы мен қалыптасуын ғалымдар АҚШ-та компьютерлік техника мен ХХ ғасырдың 60-80 жылдары британ лингвистикасындағы интеллектілік климаттың белсенді даму жағдаятымен түсіндіреді. Осы кездерде компьютерленген мәтіндік корпустарға лингвистикалық талдау жүргізу тілдік зерттеулердің ең көп бөлігін қамти бастағаны мәлім. Ғылымды институттендіру нәтижесі 2001 жылы Бирменгем университетінде корпустық лингвистика кафедрасын ашуға және International Journal of Corpus Linguistics журналын баспадан шығарып тұруға мүмкіндік туғызады. Бірнеше тілдердің материалдары бойынша құрастырылған түрі мен қызметі жағынан айырым табатын корпустар және солардың негізінде түзілген неше түрлі сөздіктер мен грамматикалар өмірге келді.

Технологиялық құралдар ретінде компьютерді қолданатын тіл білімінің басқа салаларымен қатар, корпустық лингвистика тілдік материалды жинау мен сақтаудың дәстүрлі әдіс-тәсілдеріне – жазба мәтінді қолмен өңдеу, лұғаттық картотекаларды жинау мен сақтау және т.б. түрлеріне тікелей әсер етті. Дәстүрлі тәсілдердің кемшілік жақтарына жататындар: аса көп уақыт шығыны, зерттеу нысанының көлем жағынан шектелуі, жаңартудың күрделігі, статистикалық өңдеудің қолайсыздығы, қашықтық жағдайындағы ақпарат алмасу мүмкіндігінің болмауы. Ал тілдік зерттеулерге қажетті материалдарды жинау мен сақтауға қатысты әрекеттерге компьютерлік технологияны пайдалану жаңа тәсілдің артықшылығын толығымен пайдалануға мүмкіндік тудырды. Бұл жағдайда интернет желісін пайдалану ақпарат алмасу қашықтығын жояды, қабылдайтын ақпарат көлемі талай мәрте өседі, деректерді сақтайтын және өңдейтін компьютерлік бағдарламалар арқылы қажет деген ақпаратпен танысу мен оны өңдеу әрекеттері де барынша тез жүзеге асады.

Қорыта айтқанда, нәтиженің дұрыстығын (ақиқаттығын) тексеретін теоретикалық, эмпирикалық және интуитивтік әдіс-тәсілдерге жүгіну қажет болады. Бұл жағдайда корпустық лингвистика зерттеу ақиқаттығының эмпирикалық негіздемесін қамтамасыз ететін, ал тәжірибенің жеткіліктілік принципі корпустық лингвистиканың әдіснамалық негізі ретінде саналатын тіл білімінің жеке бір бөлімі деп қарастырылады.

Корпустық лингвистика тіл білімінің жеке саласы ретінде өзімен

іргелес жатқан тіл ғылымы пәндерімен жанасып жатады, яғни математикалық лингвистика, дискурстік анализ және лексикография салаларымен жақын жатады. Корпустық лингвистиканың басқа тіл ғылымы пәндерімен қарым-қатынаста болу ерекшелігі, бір жағына алғанда, мәтіндер корпусының корпустық лингвистика қызметінің нәтижесі ретінде болса, ал екіншіден – лингвистикалық пәндердің басқа түрлеріне де бастапқы эмпирикалық материал болу мүмкіндігінде. Міне, дәл осы жағдай корпустық лингвистиканың фонетика, лексикология, грамматика және стилистика салаларымен тығыз байланыста болатындығына негіз бола алады.

Корпустық лингвистика әдістері ішінен келесі төрт топты жеке қарастырайық:

1) филологиялық әдістер;

2) теоретика-лингвистикалық әдістер;

3) математикалық (статистикалық) әдістер;

4) ақпараттық технологиялардың әдістері.

Бірінші топ әдістеріне жанрға, мәтін авторларына, жанрлар мәтіндерінің үлесіне және т.б. мәселелерге қатысты әдістер жатады. Екінші топта – мәтін таңдаудың және корпустағы лингвистикалық ақпараттың берілуінің жалпы және жеке лингвистикалық мәселелерімен байланысты әдіс-тәсілдер. Үшінші топты таңдама мәтіндердің санына, оның репрезентативтілігін (тұлғалылығын) қамтамасыз ету аспектілеріне қатысты математикалық статистиканың әдістері құрайды. Төртінші топқа корпус деректерінің компьютерлік берілісі (көрінісі) мен оны өңдеу әрекеттерін қамтамасыз ететін ақпараттану әдістері жатады.

Корпустық лингвистиканы жаңа білім аясы деп түсінуге байланысты «әдіс» ұғымының орны мен рөлі айтарлықтай өзгеріске ұшырауы мүмкін. Мәселен, корпустық лингвистиканы компьютерлік лингвистиканың бір бөлігі деп қарастыратын болсақ, онда төртінші топқа қатысты ақпараттану әдістері басымшылық етер еді. Ал В.П.Захаровтың жоғары оқу орындарында «Корпустық лингвистика» пәнін оқыту курсында мәтінге лингвистикалық шартты белгілер енгізуге, корпустарды құру технологиясына, корпустан қажетті ақпаратты іздеу әдістеріне (конкордансерілерге, корпустық менеджерлерге, сұрату тілдеріне) көбірек көңіл бөлінген [2; 5].

Мәтіндер корпусын қолдану аясының түрлілігі, олар лингвистикалық, филологиялық және педагогикалық салаларға топтастырылуы мүмкін. Ал ондай топтарға мына пәндер жатады: лексикография, грамматика, мәтін лингвистикасы, типология, контрастивтік лингвистика, аударма жасаудың теориясы мен практикасы, әдебиеттану ғылымы, лингводидактика. Қазіргі кезде корпустық лингвистиканың қолдану аясы бұрынғыдан да кеңеюде.

Корпустық лингвистиканың негізгі түсінігін, оның бейнелейтін аясына қарай, корпустағы сақталатын ақпарат бірлігі тұрғысынан, оны әртүрлі негізде топтастыру мен құрастыру және т.б. принциптерді ұстаным етуіне қарай әртүрлі ұғынуға болады. Мәселен, А.П.Барановтың тұжырымдауынша, корпустық лингвистикада негізгі назар тілдік жүйенің сөйленіс арқылы жүзеге асуына көп көңіл аударылады [1, 114-б.]. Ал В.П.Захаровтың корпус

ұғымын тар мағынада түсіндіруі қазіргі ғылыми түсінікте эволюциялық тұрғыда көрініс тауып отыр. Мысалы, ғалым «Корпусная лингвистика» атты оқу құралында былай дейді: «Под названием лингвистический, или языковой, корпус текстов понимается большой, представленный в электронном виде, унифицированный, структурированный, размечанный, филологический компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [2, 3-б.].

Корпуста сақталатын бірлік – ол пәндік саланың қайсыбір жиынтығы. Мысалы, оларға жататындар: сөз, сөзтіркес, сөйлем немесе бүтін мәтіндер жиынтығы [1]. Корпустағы деректер құрылымданған және құрылымданбаған пішімдерде сақталуы мүмкін. Біріншіде, яғни құрылымданған пішім жағдайында шартты белгілермен белгіленген мәтін туралы сөз болады. Ондай белгілер семантикалық (лексика-семантикалық), морфологиялық, синтаксистік сипатта болуы мүмкін. Мәселен, орыс тілінің ұлттық корпусында орын алған морфологиялық белгілер барлық сөздерді сөз табына, септелу категориясына, тегіне (род – муж., жен.), жіктелуіне, жанды-жансыздығына, етіс және етістік түріне, салыстырмалы шырай және т.б. морфологиялық сипаттамаларына қарай арнайы белгі қою жүргізілген. Сол сияқты семантикалық талдау арқылы да мәтін ішіне қажетті белгілердің қойылатынын айтуға болар еді. Мысалы, орыс тіліндегі үстеу сөздерге «Таксономияға», «бағалауға», «сөзжасамдыққа» қатысты белгілеулер орын алған. Мұндағы «Таксономия» белгісі: орынды, бағытты, қашықтықты, уақытты, жылдамдықты, санына және т.б. жүйелеулерге қатысты семантикалық сипаттамаларды білдіреді.

Корпустарға, сан жағынан, алғанда бірнеше талаптар қойылатыны да белгілі. Олардың негізгісі – репрезентативтілік (тұлғалылық), яғни оны сол корпустың пәндік аяның барлық қасиетін бейнелей алу мүмкіндігі деуге болады. Басқаша айтқанда, сол лингвистикалық зерттеу типіне қатысты пәндік аядағы құбылыстың кездесу жиілігі тілдік бірліктерді бір-бірінен ажырата алатындай мәнде болуы қажет [1, 118-б.]. Репрезентативтіктің бұл қасиеті сандық аспект ретінде қарастырылады да, ал оны зерттеп анықтау математикалық әдіспен жүзеге асады. Сонымен бірге репрезентативтіктің филологиялық әдіспен зерттелетін сапалық жағы да орын алады. Ал ондай сапалықты корпустағы мәтіндер жиынтығының жанрлар мен функционалдық стильдермен теңдестік сипатта болуы деп ұғыну қажет. Басқаша айтқанда, тақырыптас мәтіндер бойынша жүргізілген статистикалық зерттеу нәтижелерінің ақиқаттыққа жақындығына әрі таңдама мәтін саны, әрі олардағы авторлар саны негіз болады.

Корпусқа қойылатын басқа да талаптар бар, олар мыналар:

– релевантты құбылыстардың толық ескерілуі (тіпті пропорционалды шағындауға сәйкес келмесе де);

– корпустың үнемділік қасиеті пәндік аяны зерттеушінің де іс-әрекетін үнемдеуі;

– тілдік матералдарды құрылымдау жолының пайдаланушы үшін де аса маңыздылығы;

– корпус деректерін өңдеуге қажетті компьютерлік бағдарламамен қамтамасыз етілуі (статистикалық талдау, конкорданс құрастыру, сөздікті автоматты өңдеу – толық және ішінара сөзтізбелер алу: әліпбилі, жиілікті, лемматизациялау әрекеттері) [1].

Корпустарды келесі топтарға бөліп қарастыруға болады:

I. Құрылымдық дәрежесіне қарай: 1. Электрондық архив (сыртқы есте сақтау таспадағы өңделмеген мәтіндер); 2. Электрондық кітапхана (өңделген пішіндегі мәтіндер); 3. Мәтіндер корпусы (мәтіндердің стандартталған, өңделген және әр пәндік саланы бейнелейтін жиынтығы).

II. Тілдердің қамтылу санына байланысты:

1. Біртілді. 2. Екітілді. 3. Көптілді.

Сондықтан параллель корпустар дегеніміз – екі және көптілді корпустардың түрлері.

III. Қолдану тәсіліне қарай:

1. Зерттеу корпустары (көбінесе тілдік жүйелер қызметінің әртүрлі аспектілерін зерттеуге арналған). 2. Иллюстративтік (көрнекілік) корпустары (олар зерттеу жүргізілгеннен кейін ғана құрылады; бұл корпустардың мақсаты – алынған нәтижелерді растау және негіздеу). Мұндай қарама-қарсылық корпустардың ерте даму кезеңіне ғана тән болғаны белгілі, себебі қазіргі кездегі корпустар әрі зерттеу жүргізу үшін, әрі көрнекілік үшін де қолданылуы мүмкін.

IV. Динамикалық сипатына қарай: 1. Тілдік жүйенің белгілі қалпын бейнелейтін статикалық корпустар. 2. Динамикалық (мониторлық) корпустар. Олар пәндік аяның процессуалды аспектісіндегі тілдік феномендердің белгілі уақыт ішіндегі қызметін айқындау мәселелерін қарастырады.

V. Индекстеуге қатысты: 1. Белгіленбеген жай корпустар. 2. Белгіленген (аннотацияланған, мазмұнданған) корпустар. Қазіргі кездегі корпустар тек аннотацияланған болып келеді, себебі тек белгіленген корпустардан ғана қажетті ақпаратты іздестіруге, пайдалануға мүмкіндік бар.

VI. Тіл формасына (тұрпатына) қарай: 1. Жазба түрдегі корпустар. 2. Сөйлеу түріндегі (фонетикалық) корпустар.

Корпустық лингвистиканың жетістіктерін өзіне сақтаған аса дамыған корпус түрі – ұлттық корпус. Мұндай корпус белгілі дәрежеде ұлттық тілді толық түрде бейнелейді. Ұлттық корпустың репрезентативтілігі (тұлғалылығы) – сол тілдің жазба және сөйлеу түріндегі мәтіндерінің барлық типтерінің бейнеленуі. Ұлттық корпустың айтарлықтай дәрежеде көлемді (ондаған, жүздеген миллион сөзқолданыс) болуы репрезентативтілікке жетудің қажетті шарты болып саналады. Ұлттық корпустың ажыратылмас бөлігі оның белгіленген (аннотацияланған, мазмұндалған) бейнесі.

Ұлттық корпус ғылыми зерттеулердің түр-түрін жүргізуді қамтамасыз етеді: лексикографияға, жасанды интеллектіге, әдебиеттануға, сөйлеу тілін талдау мен жинақтауға және лингвистиканың барлық салаларына қатысты зерттеу түрлері. Сонымен бірге беделді академиялық сөздіктер мен ғылыми грамматикаларды құрастыру да корпустар негізінде жүзеге асады. Ұлттық

корпусты пайдаланушылар – әртүрлі саладағы тілшілер, әдебиеттанушылар, тарихшылар және гуманитарлық білім салаларының өкілдері. Ұлттық корпустың ана тілі мен шет тілін оқытуда, оқулықтар мен бағдарламалар құрастыруда маңыздылығы аса зор деуге болады [8; 9; 10].

Қатар тілдер (параллель) корпустары бір тілден екінші тілге аударма жасауға қатысты талдау жұмыстарын жүргізуге аса қолайлы болып келеді. Мысалы, «Орыс тілінің ұлттық корпусында» параллель мәтіндердің (қатар тілдер мәтіндері) корпустары да орын алған. Мұндай корпустар ерекше корпустар қатарына жатады. Себебі, орыс тіліндегі мәтінге оның басқа тілге аударылған үлгісі және, керісінше, шет тілдеріндегі мәтіндерге орысша аудармасы сәйкестендірілген.

Түпкі және аударма мәтіндердің бірліктері арасында «теңестіру» нәтижесінде арнайы қарастырылған шаралар бойынша сәйкестік жүзеге асады. Теңестірілген паралельді корпус – ол ғылыми зерттеулердің, әсіресе, аударма жасаудың теориясы мен практикасының аса тиімді құралы.

Жоғарыда сөз болған корпустық лингвистика пәнінің теориялық және тәжірибелік жақтары қазақ тілі мәтіндері бойынша компьютерлік корпустар базасын құру жағдайында да ескерілуі қажет. Корпустық лингвистика қазақ тіл білімінің де ерекше саласы ретінде қалыптасатын болса, қазақ тілшілеріне көлемді тәжірибелік материалдарды пайдалануға, қажетті деген тілдік деректерді тауып алуға және оларға тиісті деген өңдеулер жүргізуге мүмкіндік туындатады. Осының бәрі қазақ тіліне қатысты зерттеулердің шынайылыққа (ақиқаттыққа) жетудің эмпирикалық тәсілдеріне жаңаша көзқараспен қарауға және ғылыми айналым аясына аса маңызды тілдік материалдарды енгізуге жағдай жасайды.

Қазіргі кезде әлемдік корпустық лингвистиканың даму сипаты – ұлттық толық мәтіндерді арнайы зерттеу нысаны етіп алуды қажет етіп отыр. Сондықтан автоматтанған қазақ тіліндегі мәтіндер корпусының компьютерлік базасы (теориялық және практикалық жағынан қарастырғанда) жақын болашақта жүзеге асатын «Қазақ тілінің ұлттық корпусының» аса маңызды бастамасы болары сөзсіз. Мұндай зерттеулердің нәтижелері қазақ мәтіндерінің стилдік, құрылымдық, мағыналық, функционалдық және т.б. сипаттарын анықтауда да өзекті мәселелердің бірі болып саналады.

ӘДЕБИЕТ

1. Баранов А.Н. Корпусная лингвистика // Баранов А.Н. Введение в прикладную лингвистику: Учебное пособие. –М.: Едиториал УРСС, 2003. С. 112-137.

2. Захаров В.П. Корпусная лингвистика: Учебн.-метод. пособие. –СПб., 2005.48 с.

3. Научно-техническая информация. Серия 2. Информационные процессы и системы. 2003. №6.

4. Научно-техническая информация. Серия 2. Информационные процессы и системы. 2003. №10.

5. Захаров В.П. Корпусная лингвистика [Электронный документ] // http://download.yandex.ru/class/zakharov/CL

6. Баранов А.Н. Компьютерная лингвистика // Баранов А.Н. Введение в прикладную лингвистику: Учебное пособие. –М.: Едиториал УРСС, 2003. С. 13-38.

7. Фрэнсис У.Н. Проблемы формирования и машинного представления большого корпуса текстов // Новое в зарубежной лингвистике. Выпуск XIV. Проблемы и методы лексикографии. –М.: Прогресс, 1983. С.334-352.

8. Вербицкая Л.А., Казанский Н.Н., Касевич В.Б. Некоторые проблемы создания национального корпуса русского языка // Научно-техническая информация. Серия 2. 2003. №6. С.2-8.

9. Шаров С.А. Представительный корпус русского языка в контексте мирового опыта // Научно-техническая информация. Серия 2. Информационные процессы и системы. 2003. №6. С.9-18.

10. Национальный корпус русского языка // http://www.ruscorpora.ru.

КОРПУСТЫҚ ЛИНГВИСТИКА – ҚАЗАҚ ТІЛ БІЛІМІНІҢ ЖАҢА БАҒЫТЫ

Тілтаным (Языкознание). ҚР БжҒМ А.Байтұрсынұлы атындағы Тіл білімі институты, сәуір-мамыр-маусым. – Алматы. 2009 жыл, №2 (34) –3-11 бб.

Соңғы онжылдықта «Корпустық лингвистика» саласы ғылыми айналымға айтарлықтай дәрежеде ене бастады. Әсіресе тілдік зерттеулерде корпустарды тәжірибе жүзінде пайдалану тұрғысынан алғанда, яғни олардың сөздіктер мен неше түрлі грамматикаларды дайындау кезінде аса құнды материал болатындығы анықталды. Корпустық лингвистиканың мәселелері кейбір оқу құралдарының да арнайы тақырыбына айналып отыр [1; 2].

Сол сияқты қазақ тілінің қолданбалы саласының жаңа бағыттарының да жедел түрде дамып келе жатқанын айтуға болар еді. Қазіргі қазақ тіл білімінің даму үдерісі заман өзгерісіне сай, жаңа технологияны – компьютерлер мүмкіндіктерін ұлттық тілді зерттеу барысында кең түрде қолдануды қажет етуде. Осыған байланысты А.Байтұрсынұлы атындағы Тіл білімі институтында «Компьютерлік лингвистика» мен «Корпустық лингвистика» салаларына тікелей қатысты бар «Мәдени құндылықтар ретіндегі қазақ тіліндегі мәтіндер корпусы және сөздіктердің «Тіл – қазына» атты ұлттық компьютерлік қоры» атты тақырып бойынша ғылыми-зерттеу жұмысы жүргізіле бастады. Ал бұл зерттеу жұмысының негізгі мақсаты – қазақ тілінің мәдени құндылығы болып саналатын толық мәтіндер корпустарының мәтіндеріне қажетті деген грамматикалық белгі-кодтар енгізіп, оның дербес түрдегі «Тіл – қазына» атты компьютерлік базасын құру. Сонымен бірге мұндай компьютерлік база коммуникация қажеттігін толығымен қанағаттандыра алатындай, үлкен көлемді мәтіндер мен

сөздіктерге қатысты ақпараттық-сұраныс жүйесі түрінде көрініс табуы көзделеді. Ал толық мәтіндердің компьютерлік қорының нысандары ретінде М.Әуезовтің, Ә.Кекілбаевтың, М.Мақатаевтың, М.Мағауинның толық шығармаларынан таңдама мәтіндер алынады.

Қазақ тілінің лексикалық қорының үнемі өсіп отыру үдерісін лексикографиялық ізденістер мен баспадан шығып жатқан біртілді, екітілді сөздіктер түрлерінен де байқауға болады. Осыған дәлел ретінде, ең алғаш жарық көрген екі томдық, он томдық және қазіргі кезде құрастырылып жатқан он бес томдық «Қазақ әдеби тілі сөздіктерінің» лексикалық көлеміндерінің айтарлықтай ұлғаю сипатын айтуға болар еді.

Міне, бұл айтылғандардан қазақ тілін зерттеуде корпустық лингвистика саласына көп мән берілуі және оның әлемдік дәрежедегі теориялық және практикалық жақтарын зерттеу қажеттігі туындауда. Сонымен, біз ұсынып отырған шағын мақалада корпустық лингвистика саласының осы айтылғандардай кейбір маңызды тұстары сөз болады.

Әлем бойынша мәтін корпустарын құрастыру мен олардың қызметіне қатысты жалпы және нақты мәселелерге арналған мақалалар жарық көрген ғылымы журналдардың арнайы басылымдары да шыға бастады [3; 4]. Бірақ, әлде де, корпустық лингвистикаға қатысты көптеген мәселелер арнайы зерттеуді қажет ететіні белгілі. Оған жататындар: корпустық лингвистика мен оның негізгі ұғымдарының анықтамалары, корпустық лингвистиканың тіл білімі құрылымында алатын орны, әдіс-тәсілдері және т.б. Сонымен бірге жаңа бағыттың теориялық негізін ұғыну мәселесі корпустарды нақты зерттеулерде пайдалануға қарағанда белгілі дәрежеде қалыс қалып келе жатқаны да байқалады.

Корпустық лингвистиканы осы саланың мамандары тілдік корпустарды құру мен оны пайдалану жағдайын зерттейтін тіл білімінің бір бөлімі ретінде қарастырып келеді. Сонымен бірге, ол таңдама мәтіндер (мәтін үзінділері) түрінде көрініс табатын тілді зерттеу мен оны танып білуге қажетті маңызды материал деуге де болады. Кейбір ғалымдар ол пәннің түсінігін тар шеңберде қарастырып, оны тек компьютерлік лингвистика саласының аясында ғана түсіндіреді: «Корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с использованием компьютерных технологий» – дейді [5]. Ал компьютерлік лингвистика ұғымын, әдетте, компьютерлік құралдарды пайдаланудың кең мүмкіндігі ретінде түсіндіруге болатыны белгілі. Бұл жердегі «компьютерлік құралдар» деп отырғанымыз – компьютерлік бағдарламалар, тілдік деректерді өңдеу мен компьютерлік технологияны орынды ұйымдастыру жұмыстары және т.б. [6, 13-б.]. Ал корпустық лингвистика компьютерлерді тек «құрал» ретінде пайдаланатыны да бәрімізге аян. Міне, сондықтан да корпустық лингвистика өзіне жүктелген міндетті ондай құралсыз атқара да алмас еді. Бірақ компьютер мұндай рөлді қазіргі білім саласының барлық түрлерінде де атқаратынын ескерсек, онда олардың бәрін бірдей компьютерлік лингвистика саласына жатқыза беруге болмайды. Сондықтан біз корпустық

лингвистиканың қазіргі тілтаным парадигмасында алатын орнын анықтау үшін оның тарихына қысқаша тоқталауды жөн көрдік.

Корпустық лингвистика Н.Хомскийдің көзқарасына қарама-қарсы бағытта пайда болды деуге әбден болады. Себебі, оның оппозициялық «құзырет/түсінігі» («компетенция/представление») бойынша тіл жайындағы деректер тіл иесінің құзыретін құрайды да, сондықтан да әртүрлі аса көлемді тілдік нысанды зерттеудің орнына сөйлеушінің тілдік интуициясына (ішкі түйсігіне) сүйенгенін жеткілікті деп санайды. Корпустық лингвистика мен оған қатысты пікірталастар, айтарлықтай дәрежеде, теоретикалық және эмпирикалық білім аяларының ең ұтымды арақатынастарын іздестіру жолын бейнелейді деуге болады.

Корпустық лингвистика 1963 ж. АҚШ-та өмірге келген Браун корпусынан (The Brown Standard Corpus of American English) бастама алады. Бұл корпустың көлемі 1 млн. сөзқолданыстан тұрады және оның құрамында әрбіреуі 2 мың сөзқолданысқа тең 500 мәтін бар. Ғалымдар У.Н.Фрэнсис пен Г.Кучер бұл корпусты құрастыру үшін мынадай принциптер мен критерийлерді ұстаным еткен болатын:

1. Мәтін иесі ретінде американдық авторлармен ағылшын тілінің американдық нұсқасын қолданатын авторлар шығармаларын алу;

2. Мәтіндерді синхрондау (мезгілдік үйлестіру) үшін АҚШ-та 1961 жылы алғаш рет баспадан шыққан еңбектерді алу;

3. Жанрлардың арақатынасын таңдау жұмыстары үшін ерекше ықтималдықтар рәсімін қолдану: барлығы 15 жанр, 9 – «ақпараттық» проза, 6 – көркем проза; әрбір жанрдан 6-дан 80-ге дейін элементар таңдама мәтін алу;

4. Мәтіндерді таңдауға қойылатын талаптарды айқындау мен тұжырымдау: әр мәтіннің көлемі статистикалық тұрғыда, оның стильдік ерекшеліктерін ақиқат дәрежеде бейнелеуін ескеру.

5. Жанрлардың құрамы мен арақатынасы, олардың стильдік ерекшеліктері мен қатынастық салмағы барабарлық деңгейде көрініс табуы қажет [7].

Сонымен, Браундық корпус осыған ұқсас корпустар құруға қатысты зерттеулердің кеңінен тараған (атышулы) нысаны мен стандартына айналды. Ғалымдар арасында көптеген лингвистикалық зерттеулерді сапалы жүргізу тек ауқымды тілдік материалдар негізінде ғана жүзеге асатындығын ұғыну байқала бастады. Осы айтылғандардың барлығы мәтіндерді корпус түрінде ұйымдастыру ережелері мен оларға талдау жүргізу әдіснамасын зерттейтін бағыттың пайда болуына себепші болып, ғалымдарды одан әрі ынталандыра түсті. Сонымен корпустық лингвистика осы әдіс-тәсілдердің әдіснамасы ретінде туындап отыр деуге әбден болады.

Корпустық лингвистиканың ағылшын тілі материалында туындауы мен қалыптасуын ғалымдар АҚШ-та компьютерлік техника мен XX ғасырдың 60-80 жылдары британ лингвистикасындағы интеллектілік климаттың белсенді даму жағдаятымен түсіндіреді. Осы кездерде компьютерленген мәтіндік корпустарға лингвистикалық талдау жүргізу тілдік зерттеулердің ең көп

бөлігін қамти бастағаны мәлім. Ғылымды институттендіру нәтижесі 2001 жылы Бирменгем университетінде корпустық лингвистика кафедрасын ашуға және International Journal of Corpus Linguistics журналын баспадан шығарып тұруға мүмкіндік туғызады. Бірнеше тілдердің материалдарында құрастырылған түрі мен қызметі жағынан айырым табатын корпустар және солардың негізінде құрастырылған неше түрлі сөздіктер мен грамматикалар өмірге келді.

Технологиялық құралдар ретінде компьютерлік технологияны қолданатын тіл білімінің басқа салаларымен қатар, корпустық лингвистика тілдік материалды жинау мен сақтаудың дәстүрлі әдіс-тәсілдеріне – жазба мәтінді қол күшімен өңдеу, лұғаттық картотекаларды жинау мен сақтау және т.б. түрлеріне тікелей әсер етті. Дәстүрлі тәсілдердің кемшілік жақтарына жататындар: аса көп уақыт шығыны, зерттеу нысанының көлем жағынан шектелуі, жаңартудың күрделігі, статистикалық өңдеудің қолайсыздығы, қашықтық жағдайындағы ақпарат алмасу мүмкіндігінің болмауы. Тілдік зерттеулерге қажетті материалдарды жинау мен сақтауға қатысты компьютерлік технологияны пайдалану жаңа тәсілдің артықшылығын толығымен пайдалануға мүмкіндік тудырды. Бұл жағдайда тілдік материалға «қол жеткізу» жедел түрде іске асады, интернет желісін пайдалану ақпарат алмасу қашықтығын жояды, ақпарат көлемі талай мәрте өседі, деректерді сақтайтын және өңдейтін компьютерлік бағдарламалар арқылы қажет деген ақпаратпен танысу мен оны өңдеу әрекеттері барынша тез жүзеге асады. Сонымен бірге бұл саладағы технологиялық олқылықтар ақпарат көлемінің шексіз ұлғаюының әсерінен кейбір тіл қызметінің ақиқаттық көрінісінен ауытқуларына да әкеп соқтыруы мүмкін екендігі байқалды [1].

Жоғарыда айтылғандар «корпустық лингвистика» контексі (мәнмәтіні) мағынасында маңызды болып саналатын зерттеу нәтижесіне қатысты ақиқаттықты айқындау мәселесін қатыстыруды қажет етеді. Қорыта айтқанда, нәтиженің дұрыстығын (ақиқаттығын) тексеретін теоретикалық, эмпирикалық және интуитивтік әдіс-тәсілдерге жүгіну қажет болады. Бұл жағдайда корпустық лингвистика зерттеу ақиқаттығының эмпирикалық негіздемесін қамтамасыз ететін, ал тәжірибенің жеткіліктілік принципі корпустық лингвистиканың әдіснамалық негізі ретінде саналатын тіл білімінің жеке бір бөлімі деп қарастырылады. Корпустық лингвистиканың негізгі міндеттері ретінде мыналарды жатқызуға болады:

1. Аталған бағыттың теориялық негіздемесін жете зерттеу;
2. Корпустар түрлерін құрастыру мен қолдану тәжірибесін талдау;
3. Корпусқа қойылатын жалпы талаптарды тұжырымдау;
4. Әртүрлі зерттеулер мен оқыту міндеттерін атқаратын корпустар құрастыру;
5. Тіл білімінің әртүрлі салаларында мәтіндер корпустарын қолданудың тиімді тәсілдерін қалыптастыру. Бұл міндет зерттеуші қауымына практикалық тұрғыда аса құнды болып саналады.

Корпустық лингвистика тіл білімінің жеке саласы ретінде өзімен іргелес жатқан тіл ғылымы пәндерімен жанасып жатады, яғни

математикалық лингвистика, дискурстік анализ және лексикография салаларымен жақын жатады. Корпустық лингвистиканың басқа тіл ғылымы пәндерімен қарым-қатынаста болу ерекшелігі, бір жағына алғанда, мәтіндер корпусының корпустық лингвистика қызметінің нәтижесі ретінде болса, ал екіншіден – лингвистикалық пәндердің басқа түрлеріне де бастапқы эмпирикалық материал болу мүмкіндігінде. Міне, дәл осы жағдай корпустық лингвистиканың фонетика, лексикология, грамматика және стилистика салаларымен тығыз байланыста болатындығына негіз бола алады.

Корпустық лингвистика әдістері ішінен келесі бөліктерді жеке қарастыруға болады:

1) филологиялық әдістер;

2) теоретика-лингвистикалық әдістер;

3) математикалық (статистикалық) әдістер;

4) ақпараттық технологиялардың әдістері.

Бірінші топ әдістеріне жанрға, мәтін авторларына, жанрлар мәтіндерінің үлесіне және т.б. мәселелерге қатысты әдістер жатады. Екінші топта – мәтін таңдаудың және корпустағы лингвистикалық ақпараттың берілуінің жалпы және жеке лингвистикалық мәселелерімен байланысты әдіс-тәсілдер. Үшінші топты таңдама мәтіндердің санына, оның репрезентативтілігін (тұлғалылығын) қамтамасыз ету аспектілеріне қатысты математикалық статистиканың әдістері құрайды. Төртінші топқа корпус деректерінің компьютерлік берілісі (көрінісі) мен оны өңдеу әрекеттерін қамтамасыз ететін ақпараттану әдістері жатады.

Корпустық лингвистиканы жаңа білім аясы деп түсінуге байланысты «әдіс» ұғымының орны мен рөлі айтарлықтай өзгеріске ұшырауы мүмкін. Мәселен, корпустық лингвистиканы компьютерлік лингвистиканың бір бөлігі деп қарастыратын болсақ, онда төртінші топқа қатысты ақпараттану әдістері басымшылық етер еді. Ал В.П.Захаровтың жоғары оқу орындарында «Корпустық лингвистика» пәнін оқыту курсында мәтінге лингвистикалық шартты белгілер енгізуге, корпустарды құру технологиясына, корпустан қажетті ақпаратты іздеу әдістеріне (конкордансерілерге, корпустық менеджерлерге, сұрату тілдеріне) көбірек көңіл бөлінген [2; 5].

Мәтіндер корпусын қолдану аясы неше түрлі болғандықтан, олар лингвистикалық, филологиялық және педагогикалық салаларға топтастырылуы мүмкін. Ондай топтарға жататын пәндер: лексикография, грамматика, мәтін лингвистикасы, типология, контрастивтік лингвистика, аударма жасаудың теориясы мен практикасы, әдебиеттану ғылымы, лингводидактика. Қазіргі кезде корпустық лингвистиканың қолдану аясы бұрынғыдан да кеңейе түсті.

Корпустық лингвистиканың негізгі түсінігін, оның бейнелейтін аясына қарай, корпустағы сақталатын ақпарат бірлігі тұрғысынан, оны әртүрлі негізде топтастыру мен құрастыру және т.б. принциптерді ұстаным етуіне қарай әртүрлі ұғынуға болады.

А.П.Барановтың тұжырымдауынша, корпустық лингвистикада негізгі назар тілдік жүйенің сөйлеу арқылы жүзеге асуына көп көңіл аударылады:

«Как правило, в корпусной лингвистике языковой аспект фактически игнорируется, поскольку изначально фиксируется область привлекаемых данных – реализаций языковой системы» [1, 114-б.]. Солай бола тұра, корпусты құру мен оны түрлендіру жағын зерделеуде тілдік аспектіні ескеру қажеттігі күмән тудырмайды.

В.П.Захаровтың корпус ұғымын тар мағынада түсіндіруі қазіргі ғылыми түсінікте эволюциялық тұрғыда көрініс тапқан. Мысалы, ғалым «Корпусная лингвистика» атты оқу құралында былай дейді: «Под названием лингвистический, или языковой, корпус текстов понимается большой, представленный в электронном виде, унифицированный, структурированный, размечанный, филологический компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [2, 3-б.].

Корпуста сақталатын бірлік – ол пәндік саланың қайсыбір жиынтығы. Мысалы, оларға жататындар: сөз, сөзтіркес, сөйлем немесе бүтін мәтіндер жиынтығы [1]. Корпустағы деректер құрылымданған және құрылымданбаған пішімдерде сақталуы мүмкін. Біріншіде, яғни құрылымданған пішім жағдайында шартты белгілермен белгіленген мәтін туралы сөз болады. Ондай белгілер семантикалық (лексика-семантикалық), морфологиялық, синтаксистік сипатта болуы мүмкін. Мысалы, орыс тілінің ұлттық корпусында орын алған морфологиялық белгілер барлық сөздерді сөз табына, септелу категориясына, тегіне (род – муж., жен.), жіктелуіне, жанды-жансыздығына, етіс және етістік түріне, салыстырмалы шырай және т.б. морфологиялық сипаттамаларына қарай арнайы белгі қою жүргізілген. Сол сияқты семантикалық талдау арқылы да мәтін ішіне қажетті белгілердің қойылатынын айтуға болар еді. Мысалы, орыс тіліндегі үстеу сөздерге «Таксономияға», «бағалауға», «сөзжасамдыққа» қатысты белгілеулер орын алған. Мұндағы «Таксономия» белгісі: орынды, бағытты, қашықтықты, уақытты, жылдамдықты, санына және т.б. жүйелеулерге қатысты семантикалық сипаттамаларды білдіреді.

Корпустарға, сан жағынан, алғанда бірнеше талаптар қойылатыны да белгілі. Олардың негізгісі – репрезентативтілік (тұлғалылық), яғни оны сол корпустың пәндік аяның барлық қасиетін бейнелей алу мүмкіндігі деуге болады. Басқаша айтқанда, сол лингвистикалық зерттеу типіне қатысты пәндік аядағы құбылыстың кездесу жиілігі тілдік бірліктерді бір-бірінен ажырата алатындай мәнде болуы қажет [1, 118-б.]. Репрезентативтіктің бұл қасиеті сандық аспект ретінде қарастырылады да, ал оны зерттеп анықтау математикалық әдіспен жүзеге асады. Сонымен бірге репрезентативтіктің филологиялық әдіспен зерттелетін сапалық жағы да орын алады. Ал ондай сапалықты корпустағы мәтіндер жиынтығының жанрлар мен функционалдық стильдермен теңдестік сипатта болуы деп ұғыну қажет. Басқаша айтқанда, тақырыптас мәтіндер бойынша жүргізілген статистикалық зерттеу нәтижелерінің ақиқаттыққа жақындығына әрі таңдама мәтін саны, әрі олардағы авторлар саны негіз болады.

А.Н.Барановтың тұжырымдауынша репрезентативтікті жүзеге асырудың ең қарапайым жолы – пәндік салалар бойынша алынатын таңдама мәтіндерді сәйкес деңгейде шағындау әдісі. Міне, тек осы жағдайда ғана корпусты ұйымдастыруда сәйкестік стратегиясы сақталатын болады [1].

Корпусқа қойылатын басқа да талаптар бар, олар мыналар:

– релевантты құбылыстардың толық ескерілуі (тіпті пропорционалды шағындауға сәйкес келмесе де);

– корпустың үнемділік қасиеті пәндік аяны зерттеушінің де іс-әрекетін үнемдеуі;

– тілдік матералдарды құрылымдау жолының пайдаланушы үшін де аса маңыздылығы;

– корпус деректерін өңдеуге қажетті компьютерлік бағдарламамен қамтамасыз етілуі (статистикалық талдау, конкорданс құрастыру, сөздікті автоматты өңдеу – толық және ішінара сөзтізбелер алу: әліпбилі, жиілікті, лемматизациялау әрекеттері) [1].

Корпустарды келесі топтарға бөліп қарастыруға болады:

I. Құрылымдық дәрежесіне қарай:

1. Электрондық архив (сыртқы есте сақтау таспадағы өңделмеген мәтіндер);

2. Электрондық кітапхана (өңделген пішіндегі мәтіндер);

3. Мәтіндер корпусы (мәтіндердің стандартталған, өңделген және әр пәндік саланы бейнелейтін жиынтығы).

II. Тілдердің қамтылу санына байланысты:

1. Біртілді. 2. Екітілді. 3. Көптілді.

Сондықтан параллель корпустар дегеніміз – екі және көптілді корпустардың түрлері.

III. Қолдану тәсіліне қарай:

1. Зерттеу корпустары (көбінесе тілдік жүйелер қызметінің әртүрлі аспектілерін зерттеуге арналған).

2. Иллюстративтік (көрнекілік) корпустары (олар зерттеу жүргізілгеннен кейін ғана құрылады; бұл корпустардың мақсаты – алынған нәтижелерді растау және негіздеу). Мұндай қарама-қарсы қоюшылық корпустардың ерте даму кезеңіне ғана тән болғаны белгілі, себебі қазіргі кездегі корпустар әрі зерттеу жүргізу үшін, әрі көрнекілік үшін де қолданылуы мүмкін.

IV. Динамикалық сипатына қарай:

1. Тілдік жүйенің белгілі қалпын бейнелейтін статикалық корпустар.

2. Динамикалық (мониторлық) корпустар. Олар пәндік аяның процессуалды аспектісіндегі тілдік феномендердің белгілі уақыт ішіндегі қызметін айқындау мәселелерін қарастырады.

V. Индекстеуге қатысты:

1. Белгіленбеген жай корпустар.

2. Белгіленген (аннотацияланған, мазмұнданған) корпустар. Қазіргі кездегі корпустар тек аннотацияланған болып келеді, себебі тек белгіленген корпустардан ғана қажетті ақпаратты іздестіруге, пайдалануға мүмкіндік бар.

VI. Тіл формасына (тұрпатына) қарай:

1. Жазба түрдегі корпустар.

2. Сөйлеу түріндегі (фонетикалық) корпустар.

Корпустық лингвистиканың жетістіктерін өзіне сақтаған аса дамыған корпус түрі – ұлттық корпус. Мұндай корпус белгілі дәрежеде ұлттық тілді толық түрде бейнелейді. Ұлттық корпустың репрезентативтілігі (тұлғалылығы) – сол тілдің жазба және сөйлеу түріндегі мәтіндерінің барлық типтерінің бейнеленуі. Ұлттық корпустың айтарлықтай дәрежеде көлемді (ондаған, жүздеген миллион сөзқолданыс) болуы репрезентативтілікке жетудің қажетті шарты болып саналады. Ұлттық корпустың ажыратылмас бөлігі оның белгіленген (аннотацияланған, мазмұндалған) бейнесі.

Ұлттық корпус ғылыми зерттеулердің түр-түрін жүргізуді қамтамасыз етеді: лексикографияға, жасанды интеллектіге, әдебиеттануға, сөйлеу тілін талдау мен жинақтауға және лингвистиканың барлық салаларына қатысты зерттеу түрлері. Сонымен бірге беделді академиялық сөздіктер мен ғылыми грамматикаларды құрастыру да корпустар негізінде жүзеге асады. Ұлттық корпусты пайдаланушылар – әртүрлі саладағы тілшілер, әдебиеттанушылар, тарихшылар және гуманитарлық білім салаларының өкілдері. Ұлттық корпустың ана тілі мен шет тілін оқытуда, оқулықтар мен бағдарламалар құрастыруда маңыздылығы аса зор деуге болады [8; 9; 10].

Қатар тілдер (параллель) корпустары бір тілден екінші тілге аударма жасауға қатысты талдау жұмыстарын жүргізуге аса қолайлы болып келеді. Мысалы, «Орыс тілінің ұлттық корпусында» параллель мәтіндердің (қатар тілдер мәтіндері) корпустары да орын алған. Мұндай корпустар ерекше корпустар қатарына жатады. Себебі, орыс тіліндегі мәтінге оның басқа тілге аударылған үлгісі және, керісінше, шет тілдеріндегі мәтіндерге орысша аудармасы сәйкестендірілген.

Түпкі және аударма мәтіндердің бірліктері арасында «теңестіру» нәтижесінде арнайы қарастырылған шаралар бойынша сәйкестік жүзеге асады. Теңестірілген паралельді корпус – ол ғылыми зерттеулердің, әсіресе, аударма жасаудың теориясы мен практикасының аса тиімді құралы.

Жоғарыда сөз болған корпустық лингвистика пәнінің теориялық және тәжірибелік жақтары қазақ тілі мәтіндері бойынша компьютерлік корпустар базасын құру жағдайында да ескерілуі қажет. Корпустық лингвистика қазақ тіл білімінің де ерекше саласы ретінде қалыптасатын болса, қазақ тілшілеріне көлемді тәжірибелік материалдарды пайдалануға, қажетті деген тілдік деректерді тауып алуға және оларға тиісті деген өңдеулер жүргізуге мүмкіндік туындатады. Осының бәрі қазақ тіліне қатысты зерттеулердің шынайылыққа (ақиқаттыққа) жетудің эмпирикалық тәсілдеріне жаңаша көзқараспен қарауға және ғылыми айналым аясына аса маңызды тілдік материалдарды енгізуге жағдай жасайды.

Қазіргі кезде әлемдік корпустық лингвистиканың даму сипаты – ұлттық толық мәтіндерді арнайы зерттеу нысаны етіп алуды қажет етіп отыр. Сондықтан автоматтанған қазақ тіліндегі мәтіндер корпусының компьютерлік базасы (теориялық және практикалық жағынан қарастырғанда)

жақын болашақта жүзеге асатын «Қазақ тілінің ұлттық корпусының» аса маңызды бастамасы болары сөзсіз. Мұндай зерттеулердің нәтижелері қазақ мәтіндерінің стилдік, құрылымдық, мағыналық, функционалдық және т.б. сипаттарын анықтауда да өзекті мәселелердің бірі болып саналады.

ӘДЕБИЕТ

1. Баранов А.Н. Корпусная лингвистика // Баранов А.Н. Введение в прикладную лингвистику: Учебное пособие. –М.: Едиториал УРСС, 2003. С. 112-137.
2. Захаров В.П. Корпусная лингвистика: Учебн.-метод. пособие. –СПб., 2005.48 с.
3. Научно-техническая информация. Серия 2. Информационные процессы и системы. 2003. №6.
4. Научно-техническая информация. Серия 2. Информационные процессы и системы. 2003. №10.
5. Захаров В.П. Корпусная лингвистика [Электронный документ] // http://download.yandex.ru/class/zakharov/CL
6. Баранов А.Н. Компьютерная лингвистика // Баранов А.Н. Введение в прикладную лингвистику: Учебное пособие. –М.: Едиториал УРСС, 2003. С. 13-38.
7. Фрэнсис У.Н. Проблемы формирования и машинного представления большого корпуса текстов // Новое в зарубежной лингвистике. Выпуск XIV. Проблемы и методы лексикографии. –М.: Прогресс, 1983. С.334-352.
8. Вербицкая Л.А., Казанский Н.Н., Касевич В.Б. Некоторые проблемы создания национального корпуса русского языка // Научно-техническая информация. Серия 2. 2003. №6. С.2-8.
9. Шаров С.А. Представительный корпус русского языка в контексте мирового опыта // Научно-техническая информация. Серия 2. Информационные процессы и системы. 2003. №6. С.9-18.
10. Национальный корпус русского языка // http://www.ruscorpora.ru.

АВТОМАТТАНҒАН АҚПАРАТТЫҚ-ІЗДЕСТІРУ ЖҮЙЕСІНІҢ КЕЙБІР ҰҒЫМДАРЫ МЕН ТІЛІ

ҚР Білім және ғылым министрлігі Ғылым комитеті
А.Байтұрсынұлы атындағы Тіл білімі институты
Көрнекті тілші ғалым, белгілі түркітанушы Құлмат
Өмірәлиевтің туғанына 80 жыл толуына арналған
«Қазақ әдеби тілі тарихының өзекті мәселелері»
атты халықаралық дөңгелек үтел материалдары.
Алматы-2010, 206-214 бб.

ҚР БжҒМ А.Байтұрсынұлы атындағы Тіл білімі институтында қазақ сөзі деректерінің «Тіл – қазына» атты компьютерлік базасын құру мәселесі

жоспарлы ғылыми жұмыстар қатарында зерттелуде [1, 9-16 бб.]. Қазақ сөзінің «Тіл – қазына» атты деректер базасы жүйесінің түсінігі – көп тармақты әрі терең автоматтандырылған жүйені аңғартады және ондай тілдік модель «анатомиялық» тұрғыдан мүшеленген, ашық, бақылауға айқын, зерттеуге қолайлы әрі өзгермелі болады деп ұйғару қажет. Осындай ұстанымның негізінде қазақ сөзінің «Тіл – қазына» атты деректер базасы үлкен көлемді картотека мәтіндері мен көп көлемді сөздіктерге қатысты ақпараттық-сұраныс жүйесі түрінде көрініс табады. Аталған базаға енгізуге арналған тілдік деректер (деректер базасының нысаны): А.Байтұрсынұлы атындағы Тіл білімі институтының көп жылдардан бері жинақтаған (көлемі 5 млн.-ға жуық) картотекалық қорының мәтіні; осы мекемеде құрастырылған: «Абай тілі сөздігі», бір томдық «Қазақ тілінің сөздігі», жаңадан құрастырылып жатқан 15 томдық «Қазақ әдеби тілінің сөздігі» және басқа мекемелерде құрастырылған халық шаруашылығы, ғылым мен мәдениет салаларының көпшілігін қамтитын 31 томдық қазақша-орысша және орысша-қазаша терминологиялық сөздіктер, қазақ тілінің терминологиялық түсіндірме сөздіктері мәтіндері (11 том) [1, 29-41 бб.].

Қазіргі кезде қазақ тілінің мәтіндер корпусының электрондық нұсқасы да жаңа талаппен жоғарыда аталған «Тіл – қазына» атты электрондық деректер базасының жадына енгізілуде. Болашақ зерттеулерімізде базаға енетін ғылыми-техникалық басылымдар мен жарияланымдардың сан жағынан өсуі қажетті деген мәліметтерді тауып алу мәселесін кенеттен қиындататыны сөзсіз. Осыған байланысты біз өз жұмысымызда автоматтанған ақпараттық-іздестіру жүйесін қолдана білуіміз қажет. Ал ол үшін ең алдымен ақпараттық-іздестіру теориясын игеріп, ақпараттық-іздестіру жүйелері мен құрылғыларды құрастыру жұмысын жүзеге асыра біуіміз керек. Сондықтан да оқырман қауымына ұсынып отырған мақала қазақ тіне қатысты бұл мәселенің алғашқыларының бірі болғандықтан, оны «Автоматтанған ақпараттық-іздестіру жүйесінің кейбір ұғымдары мен тілі» деп атауды ұйғардық.

Сонымен, ең алғашқы ақпараттық-іздестіру құрылғылары «сұратуға» (запрос) сай қажетті деген құжатты тауып алу механикалық сәйкестендіру арқылы іске асыратын техникалық құрал түрінде жасалғаны мәлім. Оларда перфокарталар, суперпозициялық карталар, шеттері перфорацияланған (тесілген) перфокарталар және т.б. қолданыс тапты. Ақпараттық-іздестіру қызметінде қолданылатын есептеу техникасының дамуы *автоматтанған ақпараттық-іздестіру жүйесін* өмірге әкелді. Телекоммуникациялық құралдардың дамуы үлестірілген ақпараттық-іздестіру жүйелерін, әлемдік ақпараттық-іздестіру жүйелерін және әлемдік ақпараттық-инфрақұрылым тұтастығын жасауға өз әсерін тигізді [2; 3; 4]. Біздің баяндауымызда тек формалды түрдегі сұратулар бойынша іске асатын, қажетті мәліметті бойында сақтаған мәтінді (құжат, оның бөліктері, фактографиялық жазба) іздестіруге арналған автоматтанған ақпараттық-жүйелер тобы қарастырылады. Яғни ақпараттық-іздестіру жүйесі – мәтіндерді (құжаттарды)

немесе мәліметтерді сақтау мен іздестіруге арналған құралдар мен әдістердің жиынтығы болып табылады.

Ақпараттық-іздестірудің мәні – *құжаттың мазмұны* мен *сұрату бейнесін* өзара теңдестіру. Ал екі түрлі мәтіннің мазмұнын салыстыру дегеніміз – мәтіндердің мағыналық жақындық (ұқсастық) дәрежесін анықтау. Ақпараттық-іздестірудің негізгі ұғымы (құжат пен сұрату ұғымдарымен бірге) релеванттылық ұғым (ағыл. relevant – *қолайлы, қатысы бар*) болып табылады. Әдетте, релеванттылық ұғымын бір мәтіннің екінші мәтінге сәйкестік дәрежесі – құжат пен сұратудың сәйкестігі деуге болады. Ал релеванттылық – мағыналық релеванттылық пен сұратудың сәйкестігі ретінде де түсінік табады. Сонымен бірге релеванттылық: мағыналық релеванттылық және формалды релеванттылық түрінде де екіге ажыратылады. *Мағыналық релеванттылық* – ақпаратты сұратуға байланысты табылған құжаттың ақиқаттық сәйкестігі. Ал *формалды релеванттылық* – құжаттың іздестіру бейнесінің іздестіру ұйғарымына сәйкестігі. Бұл жердегі *іздестіру бейнесі* мен *іздестіру ұйғарымының* түсінігі дегенді құжаттың негізі мазмұны мен сұратуды арнайы ақпараттық-іздестіру тілінде *бейнелеу* деп ұғынған жөн болады. Жоғарыда аталған мағыналық және формалды релеванттылықпен бірге «пертиненттілік» (ағыл. pertinent – *қолайлы, іске қатысты*) түрін де ажыратады, яғни құжат мазмұнның шындығындағы ақпараттық-мұқтаждыққа сай келуі. Ал ақпараттық-мұқтаждық – белгілі индивидтің немесе ұйымның нақты мақсатқа жету жолындағы білімге мұқтаждығы. Ақпараттық-мұқтаждық табиғи тілдегі мәтін түрінде көрініс табады және ол *ақпараттық-сұрату* деп аталады. *Ақпараттық-сұрату* ақпараттық-мұқтаждықтың көрінісін әрқашан бірдей дәрежеде дәлме-дәл сипаттай алмайды. Себебі, ақпараттық-мұқтаждықты тұжырымды етіп айтудың субъективтік сипатта болуы әртүрлі қиындықтарға әкеледі [1, 143-148 бб.].

Релеванттылық ұғымына ақпараттық-іздестірудің негізгі нәтижелік көрсеткіші, яғни толықтық пен дәлдік сипаттары негізделеді. *Іздестірудің толықтық өлшемі* деп ақпараттық жиымда (массив) қамтылған сұранысқа сай келетін релеванттық құжаттар санының құжаттардың жалпы санына қатынасын атайды. Ал *іздестірудің дәлдік өлшемі* – берілген релеванттық құжаттар санының тауып алынған барлық құжаттар санына қатынасы.

Іздестірудің нысаны мен алға қойған мақсатына байланысты екі түрлі ақпараттық-іздестіру орын алады: құжаттық және фактографиялық. Құжаттық және фактографиялық жүйелер арасында айтарлықтай айырмашылық болмайды. Сондықтан аралас жүйелер ұғымын да атап кетуге болады. Мұндай жүйелерде құжаттық және фактографиялық ақпараттар бір ғана нысанаға қатысты болады.

Құжатты іздестіруде фактографиялық ақпарат қосымша мүмкіндік ретінде қолданылады. Фактографиялық элементтерге ие болатын құжаттық-іздестірудің бір түрі ретінде *библиографиялық-іздестіруді* атауға болады. Мұндай іздестіру тәсілі құжаттың тек библиографиялық элементтерінің баяндалуы бойынша ғана іске асады (шыққан жылы, басылым орны,

басылым түрі, авторы және т.б.). Әдетте, құжаттық-ақпарат-іздестіру жүйелерінде сақталатын мәліметтер рефератпен бірге библиографиялық көрініс беріледі. Ондай жиымдарға жасалатын *сұрату бейнесі* көбінде фактографиялық сипатта болады.

Ақпараттық-іздестірудің негізінде құжаттағы ақпарат бейнесінің сұратудағы ақпарат бейнесімен теңестіру әдісі жатқандықтан, ең алдымен олардың мазмұндарын *іздестіру жүйесіндегі* «бейне көрінісін» айқындап алу мәселесін шешу қажет болады.

Ол үшін құжат пен сұратудағы негізгі тақырыпты және олардың формалды сипаттамаларын баяндауға арналған мамандандырылған арнайы тіл – *ақпараттық-іздестіру тілі* құрастырылуы қажет [1, 146 б.].

Қазіргі кездегі автоматтандырылған ақпараттық-іздестіру жүйелерінде қолданыс табатын барлық ақпараттық-іздестіру тілін мынадай үш типке бөліп қарастыруға болады:

1) ұғымдарды жүйелі түрде топтастыруға негізделген – ақпараттық-іздестіру тілінің топтастыру типі;

2) сөз немесе сөзтіркестер арқылы көрініс табатын лексикалық бірліктердің тізбесіне негізделген – ақпараттық-іздестіру тілінің дескрипторлық типі;

3) белгілі заттық аядағы нысандар «белгі-мағына» мазмұнындағы білім моделіне негізделген – ақпараттық-іздестіру тілінің нышанды-белгілік типі.

Құжат немесе сұрату мазмұндарына көрініс беру әрекетін индекстеу деп атайды. Ақпараттық-іздестіру тілі арқылы сипаттау (жазу) кезінде құжат мазмұны қорытынды түрінде және «қысылған» пішінде көрініс табады. Ақпараттық-іздестіру тілінде жазылған құжаттың негізгі мазмұны құжаттың іздестіру бейнесі деп аталады. Ақпараттық-мұқтаждықтың табиғи тілде жазылған пішіні – *ақпараттық-сұрату* деп, ал оның ақпараттық-іздестіру тіліндегі көрінісі – *сұратуды іздестіру бейнесі* немесе *іздестіру ұйғарымы* деп аталып жүр .

Сонымен, ақпараттық-іздестіру дегеніміз – іздестіру жиымындағы іздестіру ұйғарымына сәйкес келетін құжаттың іздестіру бейнесін анықтауға әкеп соғады. Сәйкестіктің дәрежесін айқындау арнайы формалданған ережелер арқылы іске асады. Олар «шығарылу» критерийі немесе мағыналық сәйкестіктің критерийі деп аталады. Бұл жерде тек формалды ережелер мен формалды релеванттық жайлы ғана сөз болып отырғандықтан, оны мағыналық релеванттылықпен ешқашан ауыстыруға болмайды.

Шығарылу критерийінің ең қарапайым түрі ретінде құжаттың іздестіру бейнесі мен сұратудың іздестіру бейнесінің толық сәйкестігі немесе, дәлірек айтқанда, сұратудың іздестіру бейнесінің толық түрде құжаттың іздестіру бейнесіне қатысты болуы саналады. Бірақ, әдетте, іс жүзінде құжаттың іздестіру бейнесі мен сұратудың іздестіру бейнесінің жарым-жартылай сәйкестік критерийі деп те қолданыс табуда.

Шығарылу критерийіне қойылатын талап, көбіне, оның тікелей іздестіру ұйғарымына қатысты болуы, сондықтан бір ғана ақпараттық-іздестіру тілінің екі түрлі мүмкіндігі жайлы немесе бұл жерде құжаттардың

және сұрыптаудың *ақпараттық-іздестіру тілдері* туралы ғана айтуға болады. Бұл жағдайда сұрату тілі іздестіру мен шығарылу категориясына қатысты қосымша түрдегі арнайы мүмкіндіктерді өз құрамына қосып алады.

Ақпараттық-іздестіру жүйелерін олардың функционалдық белгілеріне қарай, мынадай құрама бөлшектерге (ішкі жүйелерге) бөлектеп қарастыруға болады: құжатты енгізу, сұратуларды енгізу және өңдеу, іздестіру, сөздіктерді түзу, статистиканы анықтау, іздестіру нәтижелерін өңдеу және қажетті деген құжатты ұсыну.

Аталған қызметтерді жүзеге асыратын құралдар мен әдістер *ішкі жүйелерді қамтамасыз ету* деп аталып жүр. Әдетте, *ішкі жүйелерді қамтамасыз етудің* мынадай негізгі түрлерін атауға болады: *ақпараттық қамтамасыз ету; лингвистикалық қамтамасыз ету; техникалық қамтамасыз ету; программалық (бағдарламалық) қамтамасыз ету; технологиялық қамтамасыз ету* .

Ақпараттық қамтамасыз ету – құжаттар мен сұратулардың ақпараттық жиымы және оларды сипаттаудың, құрастыру мен топтастырудың құралдары мен тәсілдері.

Лингвистикалық қамтамасыз ету – ақпараттық-іздестіру тіліндегі, индекстеу ережелері мен ұсыну критерийінен тұратын логика-семантикалық құрал.

Техникалық қамтамасыз ету – ақпаратты сақтау мен іздестіруді қамтамасыз ететін техникалық құралдар.

Бағдарламалық қамтамасыз ету – ЭЕМ арқылы ақпараттық-іздестіру жүйесінің барлық қызметін жүзеге асыратын бағдарламалық мүмкіндіктер.

Технологиялық қамтамасыз ету – автоматтанған және автоматтанбаған процестер мен ақпараттық-іздестіру жүйесінде ақпаратты өңдеу әрекеттерінің сипатталуы.

Ақпараттық-іздестіру жүйені лингвистикалық қамтамасыз ету дегеніміз – ақпаратты іздестірудің негізгі міндетін орындауға арналған логикалық және лингвистикалық құралдар мен әдістердің жиынтығы. Лингвистикалық қамтамасыз етудің құрамы, ең алдымен, тілдік құралдардан тұрады – арнайы тілдер (оның ішінде шектеулі табиғи тілді қоса есептегенде), осы тілдерге аударудың ережелері және оларды қолдану әдістері. Ақпараттық-іздестіру тілін жүзеге асыру үшін әртүрлі сөздіктер, қалыптар (формулярлар) және лингвистикалық алгоритмдер құрастырылады. Лингвистикалық қамтамасыз етудің ұғымына ақпараттық-іздестіру жүйесінде қолданылатын «шығарылу критерийі», математикалық логика мен жиындар теориясының құралдары да бірге енеді.

Лингвистикалық қамтамасыз етуге негіз болатын – *ақпараттық-іздестіру тілі*. Адамдардың қызметі жағдайында танымдық және коммуникативтік функцияларды атқаратын көптеген басқа да таңбалық жүйелерге қарағанда, ақпараттық-іздестіру тілі алғашынан-ақ формалданған ақпаратты іздестіруге арналған. Осы тұрғыдан ақпараттық-іздестіру тілінің айрықша ақпараттық-іздестіру қызметін сөз етуге болады.

Ең жақсы деген ақпараттық-іздестіру тіліне мынадай негізгі талаптар қойылады [1, 150 б.]:

1) жеткілікті түрдегі «семантикалық күштілік», яғни құжаттар мен сұратулардағы негізгі тақырыпты (тақырыптарды) толық және дәл сипаттау мүмкіндігі;

2) бірмағыналық, яғни ақпарат іздестіру тіліндегі әрбір жазбаның (запись) бір ғана мағынаға және әрбір ұғымның ақпарат іздестіру тілі арқылы біркелкі түсінікке ие болуы;

3) ашық жүйе сипатында болуы, яғни тілдің лексикалық құралдарын үнемі толықтырып, өзгертіп, түзетіп тұру мүмкіндігі.

Кез келген жазба тіл, ең аз дегенде, үш түрлі құрауыш бөліктерден тұрады: әліпби, сөздік және грамматика. Бұл жағдай ақпараттық-іздестіру тіліне қатысты да орын алады. Әртүрлі ақпараттық-іздестіру тіліндегі лексикалық бірліктерді жазуға арналған графикалық белгілер жүйесі, яғни оның әліпбиі өз құрамында латын және орыс әліпбиін, араб цифрларын, тыныс белгілерін және арнайы символдарды сақтайды. Ақпараттық-іздестіру тілінің тілдік бірліктерінің жиынтығы *сөздік* деп те аталады және оған тілдік бірлік ретіндегі ұғымдар мен ұғым топтары да бірге саналады.

Тілдік бірлік араларындағы қатынастарды анықтайтын жиынды *ақпараттық-іздестіру тілінің грамматикасы* деп атайды. Бұл қатынастар екі түрлі болып кездеседі: синтагматикалық (мәтіндік) және парадигматикалық (базалық). *Синтагматикалық қатынастар* дегеніміз – нақты мәнмәтінде (яғни осы құжатта немесе сұратуда) белгілі бір жағдаятта пайда болатын тілдік бірліктердің араларындағы қатынастар.

Парадигматикалық қатынастардың туындауы үшін ақиқат әлемінің нысандары мен құбылыстарын бейнелендіретін элементтердің орнына жүретін ұғымдар арасындағы логикалық байланыстың шарттылығы алынады. Табиғи тілдегі шағын ұғымдық түсінікте, грамматика дегеніміз қатынастардың бірінші типіне ғана қатысы бар мәселелер. Синтагматикалық қатынастарды белгілей алатын құралдары бар ақпараттық-іздестіру тілін «грамматикалық-ақпарат іздестіру тілі», ал мұндай құралдары жоқ ақпараттық-іздестіру тілін «грамматикасыз ақпараттық-іздестіру тілі» деп атайды. Парадигматикалық қатынастар тілдің сөздік құрамына қатысты болады. Табиғи тілдердегі сөздіктерге қарағанда, ақпараттық-іздестіру тіліндегі сөздікте парадигматикалық қатынастар нақты түрде көрініс табады.

Аталған жағдайларды ескере отырып, ақпараттық-іздестіру тілінің жалпы топтастыруын мынадай түрде беруге болады [1; 4; 5]:

1. Алдын ала үйлестірілген ақпараттық-іздестіру тілі.

1.1. Иерархиялық топтастырулар.

1.2. Әліпбилі-пәндік топтастырулар.

1.3. Фасетті (біржақты) топтастырулар.

2. Соңынан үйлестірілген ақпараттық-іздестіру тілі.

2.1. Дескрипторлық тілдер.

2.2. Семантикалық тілдер.

2.3. Синтагматикалық тілдер.

Алдын ала үйлестірілген ақпараттық-іздестіру тілінде барлық кластардың аттары да алдын ала беріледі. Әрбір атауға қатысты табиғи тілдегі оған сәйкес келетін ұғымның атауы да беріледі .

Автоматтандырылған ақпараттық-іздестіру жүйелердің даму сипаты, негізінен, *дескрипторлы типті ақпараттық-іздестіру тіліне* қатысты болады.

Дескрипторлық тілдерді құрастыруға координаттық индекстеу принципі негіз етіп алынады. Бұл ұстаным бойынша құжаттың негізгі тақырыбы – дескрипторлардың тізбесі, яғни қарапайым ұғымдардың аттарын білдіретін сөздер арқылы сипатталады.

Құжаттар мен сұратуларды тірек сөздер, яғни индекстеуге қажетті мәтіннен алынған терминдер арқылы индекстеу кезінде ақпараттық-іздестіру тілін лексикалық бақылау қажеттігі туындайды. Лексикалық бақылау – координаттық іздестіру кезіндегі сөздердің көпмағыналығы мен біркелкі қолданбауының теріс әсерін барынша азайту мақсатында қолданылатын шаралар жиынтығы. Лексикалық бақылаудың негізгі мақсаты – мәтіндегі берілу тәсіліне қарамай, құжат пен сұрыптаудың іздестіру бейнелері ұғымдарының бірмағыналы түсінігін қамтамасыз ету. Бұл жердегі лексикалық бақылау термині *сөздіктік бақылау* сөз тіркесімен синонимдес. Сондықтан ақпараттық-іздестіру тілінде лингвистикалық бірліктерді нормаға келтіру үшін барлық маңызды тірек сөздер мен дескрипторлардың тізімі, ашық түрде берілген *индекстеу сөздігі* немесе *дескрипторлық сөздік* қолданылады. Парадигматикалық қатынастары қамтылған дескрипторлық сөздік – *ақпараттық-іздестіру тезаурусы* (грек thesauros – *қор*) деп аталады.

Тезаурус арқылы іске асатын лексикалық бақылаудың мәні негізінен мына төмендегіге саяды:

1) синонимдік сипатты бақылау;

2) омонимдік сипаттың басын ашу;

3) бір сөзге ғана қатысты оның барлық түрлерін (варианттарын) нормалау, яғни олардың бәрінің басын біріктіру.

Енді лексикалық бақылауды құраушылардың әрбіреуіне қысқаша тоқталайық [1; 5; 6]:

1. Синонимдерді бақылау әрекеті ақпараттық-іздестіру жүйесіндегі лексикалық бақылаудың ең негізгі қызметі деуге болады. Мұндай әрекеттің барлық синонимдері тек бір ғана лексикалық вариантпен ауыстырылады. Бұл вариант *дескриптор* деп аталады және дескрипторлық тілдің лексикалық бірлігі орнына жүреді. Дескриптор жиынтығы және оған мағыналық жағынан жақын жатқан сөздер шартты эквиваленттіліктің тобын құрайды. Синонимдер қатарына мыналар қосылады:

- толық синонимдер (радар – радиолакатор, ЭВМ – компьютер);
- сөздердің фонетикалық (графикалық) варианттары;
- аббревиатуралар (ақпарат іздестіру жүйесі);
- мән-мағыналық (смысл) жағынан бір ғана мағынасымен сәйкес келетін сөздер (компьютер – машина, мақала – жұмыс).

2. Лексикалық бақылау омонимдерді (омографтарды) және көпмағыналы сөздерді ажыратуды да өз міндетіне алады. Мұндай мақсат оларға әртүрлі кодтарды (шартты белгілерді) сәйкестендіру арқылы іске асады. Ақпарат іздестіру тілінде лексикалық бірлік сандық (цифрлық) әліпби арқылы жазылады немесе таңбалар (белгілер) жүйесі арқылы беріледі. Кейбір зерттеулердің нәтижелері бойынша омонимдерді ажыратпау ақпараттық жағынан көп кедергі келтірмейді.

3. Мәтінде кездесетін сөзтұлғалар тірек сөздер рөлін атқарып, индекстеу процесінде осы тірек сөздің (дескриптордың) сөздік формадағы нормаланған түріне (мысалы, атау септігі, жекеше) ауыстырылады. Кейде шартты эквиваленттіліктің бір ғана класына бір лексеманың грамматикалық формасына сайып қоймай, сонымен бірге әртүрлі негізді, тіпті әртүрлі сөз таптарына қатысты лексемалар ортақ лексикалық мағына негізінде (гиперлексема) біріктіріледі. Мысалы, сверло (тескіш) – сверление (тесу) және т.б.

Тезаурустың лексикалық бақылаудан кейінгі екінші маңызды қызметі – ақпараттық-іздестіру жүйесінің нәтижелігін, яғни толықтығын және дәлдігін арттыру. Бұған тек парадигматикалық және синтагматикалық қатынастарды қолдану арқылы қол жеткізуге болады [1-3; 5].

Сонымен, ақпараттың әрбір түрі өз мазмұны арқылы сипатталады да, бірақ олардың бәрі бірдей машиналық (компьютерлік) көрініске қатысты жалпы заңдылыққа бағынады. Сондықтан информатика ғылымының негізгі міндеті ақпараттың барлық түрін қазіргі техникалық құралдар (компьютерлер) арқылы дайындау, өзгерістер енгізу, ұзақ мерзімге сақтау, тұтынушы қауымға тарату мен оны қолданудың заңдылықтарын зерттеу [6; 7].

Әдебиет

1. *Жұбанов Асқар Құдайбергенұлы.* Қазақ сөзінің **«Тіл – қазына»** деректер базасы және оның теориялық негіздері. – Алматы: «Арыс» баспасы, 2009. – 304 бет.

2. *Черный А.И.* Введение в теорию информационного поиска. М., 1975.

3. *Ланкастер Ф.* Информационно-пойсковые системы. М., 1972.

4. *Панков И.П., Захаров В.П..* «Информационно-поисковые системы» / В учебнике «Прикладное языкознание», отв А.С.Герд, СПб, 1996.

5. *Москович В.А*. Информационные языки. М., 1971.

6. Информатика / Под ред. К.В.Тараканова. М., 1986.

7. *Дейт К.* Введение в системы без данных. М., 1980.

ҚАЗАҚ ТІЛІНІҢ «ТІЛ – ҚАЗЫНА» АТТЫ КОМПЬЮТЕРЛІК БАЗАСЫ ЖАЙЫНДА

атты халықаралық ғылыми-теориялық конференция. – Астана, 2010. Л.Н. Гумилев атындағы Еуразия ұлттық университеті. -298-304 бб.

Ғылыми-техникалық прогресс, электронды-есептеу техникасының дамуы, нақты ғылымдар салаларының әдіс-тәсілдерінің түрлі білімдерде қолданыс табуы түркі тілдеріне, соның ішінде қазақ тіліне де өз ықпалын тигізуде.

1969 жылы Қазақстан Ғылым академиясы Тіл білімі институтында «Статистическое и информационное изучение тюркских языков» атты Бүкілодақтық ғылыми жиынның өтуі қазақ тіл білімі үшін елеулі құбылыс болды. Осы кезеңнен қазақ тілін статистикалық зерттеу мен автоматтандыру әдістері жүйелі сипатқа ие бола бастады.

«М.О.Әуезовтің «Абай жолы» романының жиілік сөздіктері» және «М.О.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері» атты еңбектерді қазақ тілі мәтіндерін статистика тәсілімен зерттеудің және лексикография саласын автоматтандырудың алғашқы нәтижелері деп қарастыруға болады [1; 2]. Жазушы тілінің лексикалық байлығын, автордың сөз қолданудағы стильдік тәсілінің өзіндік ерекшелігін, қазақ тілінің көркемдік және бейнелеуіш мүмкіндіктерін меңгеру деңгейін осы аталған лексикографиялық еңбектер бойынша бағалауға мүмкіндік бар.

Электронды есептеу машиналарының (ЭЕМ) көмегімен ұлттық тілдерді зерттеу мен одан туындайтын ақпараттарды автоматты түрде танып-білудің негізінде XX ғасырдың 80-жылдарынан бастап тілдердің машиналық қорын (базасын) жасау жұмысы жайлы ой-пікірлер жүзеге аса бастаған болатын. Осыған байланысты 1983 жылы Мәскеуде шақырылған Бүкілодақтық ғылыми конференцияның күн тәртібіне орыс тілінің машиналық қорын жасау проблемасы алғаш рет қойылды. Бұл ғылыми мәжілісте ЭЕМ-ның жадына түсірілетін тілдік қордың құрамы мен құрылымы қандай болуы тиіс екендігі, бірінші кезекте мұндай қор қандай міндеттер атқаратыны, тіл көздерінің ауқымы қандай болмақтығы жайлы және т.б. көптеген мәселелер қаралған болатын. Сонымен бірге электрондық тіл қорын дәйекті түрде тұтыну үшін қандай ұйымдық және техникалық шаралардың атқарылу қажеттігі де ортаға салынып, жан-жақты талқыланды [3].

Аталған ғылыми конференция шешімдерінің орындалу жағдайын қадағалау мен жаңадан туындайтын мәселелерді шешу қажеттігінен 1987 жылы Мәскеуде ашылған Бүкілодақтық екінші ғылыми конференцияда осы саладағы ғалымдар тағы да бас қосты. Бұл ғылыми форумға Қазақстанның және басқа да одақтас республикалардың ғалымдары қатысты.

Мәжілісте өткен төрт жыл аралығында орындалған жұмыстардың нәтижелері сөз болуымен қатар, басқа да ұлт тілдерінің машиналық қорларын жасау мүмкіншіліктеріне байланысты соны мәселелер қарастырылды. Ерекше атап кетерлік жайт – осы ғылыми жиында орыс, украин, грузин, эстон тілдеріне қатысты баяндамалардың қатарында қазақ тілінің машиналық

қорын жасау мәселесі де өз алдына тыңдалды (баяндамашылар – Қ.Бектаев және А.Қ.Жұбанов).

Сөйтіп, орыс тілінің машиналық қорын жасау мәселесіне арналған екінші ғылыми конференцияның шешімдерінде қазақ мәтіндерін ЭЕМ жадына енгізу және оны автоматты жолмен статистикалық тәсілдермен зерттеу бағытындағы қазақ ғалымдарының қол жеткізген нәтижелері ескеріліп, түркі тілдерінің машиналық қорын жасайтын координациялық орталық Қазақ ССР ҒА Тіл білімі институтының құрамынан ашылсын деген дербес қаулы қабылданғаны біз үшін үлкен абырой болды.

Көп ұзамай, 1988 жылдың мамыр айында Мәскеуде Кеңес Одағы түркітанушы комитетінің XIV Пленумы ашылып, оның күн тәртібіне түркі тілдерінің машиналық қорын жасау мәселесі арнайы қойылды. Бұл ғылыми жиында негізгі баяндамашылар ретінде Одақ бойынша белгілі ғалымдар: Р.Г.Пиотровский, А.М.Щербак, В.Г.Гузев өз ой-пікірлерін ортаға салып, ондай қор жасаудың қажеттігімен қатар алда тұрған қиыншылықтарын да атап өтті [4]. Ал қазақ тілінің машиналық қорын жасау мүмкіндігі жайлы мәселеге байланысты қазақстандық ғалымдар атынан Қ.Б.Бектаев және оның шәкірті А.Қ.Жұбанов баяндама жасаған болатын.

Аталған Пленумда компьютерлік қор жасаудың методологиялық негіздері ретінде мынадай екі тезисті басшылыққа алу ұсынылды: тіл жөніндегі деректердің қандайы болмасын лексикографиялық пішінде болуы және лексикографиялық жағынан жүйеленген тілдік мәліметтердің барлығы бірдей компьютерлік алгоритмдер тіліне лайықталуы.

Тілдік деректердің лексикографиялық пішінде көрініс табуы олардың сөздік түріндегі жұмыстардың идеясына жақындайды, ал мұндай тұжырым тілдік ақпараттарды шоғырландыратын бірегей түйін немесе түпқазық – жеке сөз деп ұйғарумен бірдей. Шынында да, сөз өзінің ішкі-сыртқы қызметтері жағынан және ақпарат сақтағыштық қасиеттерінің негізінде белгілі бір жүйедегі сөз қорын танытатыны сөзсіз. Сондықтан сөздің белгілі мөлшерде өзіндік сегменттері орын алады және олар грамматикалық, стилистикалық, фонетикалық, контекстік т.б. пішінде көрініс табады, яғни ол барлық тілдік мәліметтер жүйесін сәулелендіреді деп ұйғару қажет.

Көптеген ғалымдардың пікірлерін қорыта келе, біз қазақ сөзінің «Тіл – қазына» атты деректер базасы жүйесінің түсінігі – көп тармақты әрі терең автоматтандырылған жүйені аңғартады демекпіз және оның атқаратын міндеті ақпараттық барлау мен тіл білімі саласындағы (жалпы филология бағыттағы) зерттеушілік болуы тиіс. Мұндай «Тіл – қазына» атты формалды компьютерлік жүйе тілдің жаратылысына сәйкес әрі тең көлемді болуы қажет, сөйте тұра ондай тілдік модель «анатомиялық» тұрғыдан мүшеленген, ашық, бақылауға айқын, зерттеуге қолайлы әрі өзгермелі болуы тиіс деп ұйғарған жөн.

Осындай ұстанымның негізінде ҚР Білім және ғылым министрлігі А.Байтұрсынұлы атындағы Тіл білімі институтында қазақ сөзі деректерінің «Тіл – қазына» атты компьютерлік базасын құру мәселесіне байланысты біршама ғылыми жұмыстар орындалды. Бұл ретте жоспарлы тақырыптар

бойынша түзіліп жатқан түсіндірме, екі тілдік, бір тілдік сөздіктер мәтіндерінің Институттың «сөз қоржынында» компьютер жадында электрондық нұсқада тұрақты сақталуын, дәстүрлі және қазақ тілінің компьютерлік лингвистика бағыты бойынша жүргізіліп жатқан басқа да қолданбалы зерттеу жұмыстарын атап айтуға болады.

Міне, сол себепті қазақ тілінің лексикалық қорын ашық жүйе ретінде қарастыра отырып, қазақ сөзінің «Тіл – қазына» атты деректер базасын құру мәселесі ҚР БжҒМ А.Байтұрсынұлы атындағы Тіл білімі институтының кезекті зерттеулерінің күн тәртібіне енді [5]. Бұл базаға алғашқы енгізілетін нысан ретінде көп жылдық тарихы бар және аса ұқыптылықпен жинақталған көлемі, шамамен алғанда, 5 миллиондай болатын қазақ тілінің картотекалық қоры алынды.

Ал осындай деректер бойынша компьютерлік база құру мақсатында мына міндеттер жүзеге асырылуда:

1) компьютерлік база құрудың әлемдік тәжірибесі мен алдын ала белгіленген зерттеу тақырыбының талаптарына сай, аталған картотека мәтінін электрондық пішінге келтіру;

2) компьютерлік лексикография саласының «тілдік деректер» негізінде компьютерлік база құру жолындағы объективті деректер базасын, объективті-реляционды деректер базасын, СУБД мен Web және кешенді талдау құралдарын (OLAP) қамтитын әлемдік тәжірибедегі жаңа технологиялар негізінде компьютерлік алгоритмін құру және оның жасанды тілдегі бағдарламасын жазу;

3) қазақ тілінің картотекалары мен сөздіктерінің «Тіл – қазына» атты деректер базасының жобасын үш түрлі әдістеменің ұштасуы арқылы, яғни концептуалды жобалау, логикалық жобалау және физикалық жобалау әдістемелері арқылы жүзеге асыру;

4) қазақ тілінің картотекалық және сөздіктерінің «Тіл – қазына» атты деректер базасын әрі бөлек пішінде, әрі олардың объективті-реляциялық база арқылы бір-бірімен тығыз байланыста болуын ескеру;

5) қазақ сөзінің картотекалық «Тіл – қазына» атты деректер базасы «деректер базасын басқару жүйесі» (ДББЖ) арқылы іске қосылып, тұтынушының қажетті сөзге байланысты мәтіндік қоршауымен жұмыс атқаруының барлық талаптарын қанағаттандыру мүмкіндігін қарастыру;

6) қазақ сөзінің «Тіл – қазына» атты деректер базасы үлкен көлемді картотека мәтіндері мен көп көлемді сөздіктерге қатысты ақпараттық-сұраныс жүйесі түрінде көрініс табуы және аталған деректер базасының нысаны ретінде: А.Байтұрсынұлы атындағы Тіл білімі институтының көп жылдардан бері жинақтаған картотекалық қоры мәтінін, осы мекемеде құрастырылған: «Абай тілі сөздігі», бір томдық «Қазақ тілінің сөздігі», жаңадан баспадан шыға бастаған 15 томдық «Қазақ әдеби тілінің сөздігі» және басқа мекемелерде құрастырылған халық шаруашылығы, ғылым мен мәдениет салаларының көпшілігін қамтитын қазақша-орысша және орысша-қазаша терминологиялық сөздіктері, қазақ тілінің терминологиялық

түсіндірме сөздіктері мәтіндері базаға енгізуге арналған тілдік деректер [5, 29-41 б.].

Формалды пішіндегі қазақ сөзінің «Тіл – қазына» деректер базасынан күтілетін ең маңызды мәселелерге жататындар:

а) тұтынушыға қажет болған белгілі бір сөзді қандай тіл көздерінен, мәтіндерден, тіл айырымдарынан іздестіру қажет;

ә) іздеуге тиісті сөз қарастыратын мәтінде қамтылған ба, жоқ па?

б) графикалық және фонетикалық пішіндерін қоса есептегенде сөздің мүмкін болған барлық түрлері мен тұлғасы айқындалып болды ма?

в) белгілі бір сөзге қатысты мәтін үзінділерінің (контекстердің) жалпы аумағы, сипаты анықталды ма?

г) белгілі бір сөз: қарастырып отырған семантикалық және грамматикалық сипатта өткен дәуірлердегі, көне ғасырдағы мәтіндерде қолданылған ба, егер қолданылса, олар қандай сөздіктерден көрініс тапқан?

Кез келген тілдің сөздік құрамы жүздеп-мыңдап саналады. Сондықтан қажетті сөзді тез іздеп табу үшін лексиканың жүйелік сипатын ескеретін стратегияны туындатып алған жөн болады. Соның негізінде қажетті деген сөзді барлық сөздік құрамның ішінен іздемей, тек оның кішігірім бөлігінен, мысалы, семантикалық өріс ішінен немесе лексика-семантикалық топ ішінен және т.б. іздеу мүмкіндігі туындайды. Оқу лексикографиясының теориясы, ең ұтымды жағдайда, тіл жүйесінің лексикалық аспектісі бойынша, аспектілік сөздіктер жүйесін құруды қалайды. Сол арқылы іздеудің табиғи стратегиясы модельденіп, қажетті сөзді іздеп-табу уақыты қысқарады.

Бұл мәселе сөздіктің көлеміне, оларды ұйымдастыруға, сөздіктегі әрбір лексикалық бірлікті сақтайтын ақпарат көлеміне, әрқилы таңбалар жүйесіне, сілтемелер жүйесін құру жұмыстарына да қатысы бар. Яғни бұл мәселе кешенді сипатқа ие болу керек деген сөз.

Қазіргі кезде халықаралық қатынас нығайып, көптеген дербес ұлттық мемлекеттердің пайда болуы және бұрынғы отарлықта қалған халықтардың мәдениетінің көтерілу сипаты, қайтадан лексикографияның тіларалық қарым-қатынасқа және тілдерді оқытуға байланысты қызметтерін бірінші орынға ауыстыруына жағдай жасауда. Әлемдік өндірісті интеграциялау мәселесінде әртүрлі түсініктерді бірізге түсіру ісі де аса қажет болуда. Сондықтан егемен Қазақстанның халықаралық деңгейге шығуына және бірқатар жетекші елдермен қазіргі таңдағы ғылым мен технологияның әртүрлі салаларында ғылыми және өндірістік ақпараттармен алмасудың едәуір ұлғаюына байланысты неғұрлым кең ауқымды қамтитын көптілді сөздіктер шығару қажеттілігі де туындауда. Жақын болашақтағы лексикографиялық зерттеулердің негізінде оқу сөздіктерінің әр ұлттық тілдерге, соның ішінде қазақ тіліне қатысты да практикалық маңызы зор жүйесі құрылуы мүмкін.

Қолданыстағы терминологиялық деректер қорының көпшілігі ғылыми-техникалық әдебиетті аудару үшін құрылып жатқандығы белгілі. Сондықтан қазіргі кезде қазақ сөзінің «Тіл – қазына» атты деректер базасын терминологиялық деректермен де қызмет етуін алдын ала қарастырудың да маңыздылығы күннен-күнге артуда.

Белгіленімнің әрбір типі пайдаланушылардың нақты бір санатына бағытталуын және терминологиялық деректердің белгілі бір элементтерінің болуын талап етеді. Мысалы, анықтамалық терминологиялық қызметке бейімделген сауалдардың негізгі бөлігі терминдер мағыналарының шегін, олардың қолданылу мәртебесі мен сипатын анықтаумен, сонымен қатар терминдердің қолданылуын ретке келтіретін қолданыстағы нормативтік құжаттарды және т.б. іздестірумен байланысты. Көпшілік жағдайда аудармашылар үшін қолданылу ерекшеліктері жайлы ақпараты бар шеттілдік терминдердің дәл баламасы қажет болуы мүмкін. Ал **мәтінді автоматты өңдеу** (МАӨ) жүйесі мәтіндерді морфологиялық, сөзжасамдық және синтаксистік талдау мен жинақтау үшін терминдер туралы грамматикалық ақпараттың болуын талап етеді. Терминдерді ақпараттық-іздестіру тілдерінің лексикалық бірліктері ретінде пайдалану синонимдік, гипонимдік (тектік-түрлік), партитивтік және ассоциативтік байланыстар, сол сияқты ақпараттық құжаттарда терминдердің жиі, сирек қолданылуы жөніндегі мәліметтердің болуын қажет етеді. Терминологиялық жұмыстар, бірінші кезекте, термин түзушілерді қарастырып отырған саладағы және іргелес пәндік салалардағы бірізге келтірілген терминдер туралы деректермен қамтамасыз етулері қажет. Сан салалы лексикографиялық материалдарды дайындау және лингвистикалық, терминологиялық зерттеулерді жүргізу үшін терминдердің фонетикалық және графемалық сипаттамасын, қолданылу саласын, олардың жасалу тәсілдері мен үлгілері туралы деректерді қоса алғанда, жан-жақты, алуан түрлі мәліметтер де аса қажет.

Барлық сөздіктер (ал терминологиялық деректер базасын айрықша типтегі автоматтандырылған сөздік ретінде қарастырған жөн) негізгі екі қызметті атқарады – тізбелеу және нормативтік (дескриптивтік және прескриптивтік). **Терминологиялық деректер базасын** (ТДБ) ұйымдастыру мен терминологиялық ақпаратты жинаудың бірінші кезеңінде тізбелеу мәселесінде қолданыстағы терминология мен ол туралы әртүрлі дереккөздерде тіркелген ақпаратты барынша сипаттап беру мәселесі елеулі орын алады.

Қазақ сөзінің «Тіл – қазына» атты терминологиялық деректер базасының мақсаты адам-компьютер жүйесіндегі ғылыми-техникалық сөйлеу актілерінде тілдік норманы жүзеге асыру болып табылады. Алайда тілдік норманы жүзеге асыру ғылыми-техникалық тілдің барлық элементтерін тізбелеу негізінде осы норманы алдын ала анықтап алмайынша мүмкін болмайды.

Қазақ сөзінің «Тіл – қазына» терминологиялық деректер базасының тізбелеу қызметі мынаған бағытталған:

– терминологиялық лексиканы барынша толық қамту және жан-жақты сипаттау;

– терминологиялық лексиканың барлық қолданылу аспектілерін көрсететін алуан түрлі дереккөздерді барынша енгізу;

– терминологиялық жүйелерді құрастыру мен дамытуға ынталы барлық ұйымдармен қарым-қатынас жасау.

Қазақ сөзінің «Тіл – қазына» атты терминологиялық деректер базасынның тізбелеу қызметі тек ғылыми ынталылықпен ғана емес, сонымен қатар ақпараттық орталықтарды, баспаларды және басқа да ұйымдарды өздері жасаған тілдік құралдарды ұсынуға міндеттейтін ұйымдастырушылық шешімдермен де нығайтылуы керек.

Қазақ сөзінің «Тіл – қазына» атты деректер базасының нормативтік қызметі ең алдымен мынадан көрінеді, ол арнаулы тілдік стильдерді банктегі деректерге сәйкес лексикалық құрамға енгізе отырып, олардың практикалық тұрғыдан жасалуына қатысады, қазақ тілінің машиналық қоры бірыңғай, бірегей және жалпымемлекеттік тілдік орган ретінде құрылымданатын болады.

Ғылыми-техникалық «адам-компьютер» қарым-қатынасы саласындағы тілдің қолданылу аясын анықтай отырып, қазақ сөзінің «Тіл – қазына» терминологиялық деректер базасы тілдің барлық стильдері мен салаларына жаппай ықпал етеді. Бұл ықпал жалпытілдік норманы бекіту мен дамыту бағытында болуы тиіс. Сондықтан да нормативтік қызмет саналы түрде терминологиялық деректер банкін жасаудың мақсаты ретінде қойылады. Оның аясына қазіргі таңда жеке-жеке қолданылатын төрт нормалаушы (қалыптандырушы) ағым біріктірілген:

– мемстандарт басылымдарында бекітілген терминдер мен анықтамалар стандарттары бойынша ғылыми-техникалық терминологияны стандарттау;

– мемтерминком бюллетеньдерінде көрсетілген ұсыныстар бойынша ғылыми терминологияны нормалау;

– сәтті терминдер (дескриптролар) мен сәтсіз терминдерді (аскрипторлар) бөліп алуға қатысты ақпараттық-іздеу тілінің лексикалық бірліктерін ажырату;

– белгілі бір терминді қолдану жөніндегі эпизодтық жарияланымдар мен ұсыныстарда көрсетілгендей, тіл мәдениеті мамандарының жеке күш-жігерін салуы.

Осы аталғандарды негізге алған бағыттар терминологиялық деректер базасына енгізілген әр терминнің арнайы сипаттамаларында беріледі.

Болашақта қазақ сөзінің «Тіл – қазына» атты деректер базасы адам-компьютер қарым-қатынасының барлық түрін терминологиялық ақпаратпен қамтамасыз ететін болады. Бұл ретте тәжірибе жүзінде қолдану үшін ғылыми-техникалық терминдерді семантикалаудың компьютерлік жүйесін жасау және енгізу міндеті алдыңғы орынға шығады.

Міне, осы аталғандар қазақ сөзінің «Тіл – қазына» атты деректер базасын құру мақсатындағы күн тәртібіне қойылатын ең маңызды мәселелер деуге болады.

Тілдің машиналық қорының қажеттілігін алғаш дәлелдеп берген ғалым – информатика саласының белгілі маманы А.П.Ершов тілдің машиналық қорын құрастыру мәселесіне үлкен мән берген болатын [6]. Оған ғалымның 1978 жылы айтқан мына сөздері айғақ: «Орыс тілін сүйетін қоғамның (формалды қоғам) танымынша, орыс тіл білімі дегеніміз – әзірге ұйықтап

жатқан Гулливер тәрізді, сондықтан менің ойымша, оның оянатын және өзі туралы бар дауыспен жар салатын кезі енді жетті. Ал ЭЕМ бұл ретте таптырмайтын қоңыраулы сағат секілді» [3, 7-12 б.]. Бұл айтылғандар қазақ тілінің қазіргі жағдайының компьютерлік лингвистика саласына қатысты әлі де өз күшін жоғалтпаған тәрізді. Ал осы саланың белгілі ғалымы А.П.Ершов сол кездің өзінде-ақ аса білгірлікпен болжап, болашақ жасалатын орыс тілінің машиналық қорын тіл жөніндегі білімдердің аса қуатты концентраты (қоспасы) болады деп сенген.

Толық сипаттағы қазақ сөзінің «Тіл – қазына» деректер базасының ішінде неше түрлі мәтіндер, сөзтұлғалардың статистикасы, сөзнұсқағыштардың толық тізімдері, сөздерге телулі цитаталар, жиілікті және кері-әліпбилі сөздіктер көрініс табатыны сөзсіз. Аталған тақырып өзегіне қатысты барлық еңбектер, ғылыми зерттеу жұмыстары бір ізге, бір жүйеге түсіп, зор жетістіктерге жетеді деп білеміз.

Сонымен, қазақ сөзінің «Тіл – қазына» деректер базасының теориялық және практикалық маңыздылығы қазақ тілінің болашағы үшін аса жоғары демекпіз. Сондықтан да қазақ сөзінің «Тіл – қазына» атты көпаспектілі компьютерлік базасын ғылыми тұрғыда құру – қазақ лексикографтары үшін жаңаша ойлап, жаңаша шешімге келудің ең тиімді жолы деуге болады.

Әдебиет

1. *Бектаев Қ.Б., Жұбанов А.Х., Мырзабеков С., Белботаев А.Б.* «М.О.Әуезовтің «Абай жолы» романының жиілік сөздігі». – Алматы: Ғылым, 1979. -336 б.

2. *Бектаев Қ.Б., Жұбанов А.Х., Мырзабеков С., Белботаев А.Б.* «М.О.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері». – Алматы-Түркістан: 1995. – 346 б.

3. Машинный фонд русского языка: идеи и суждения. М.: Наука, 1986. 240 с.

4. *Гузев В.Г., Пиотровский Р.Г., Щербак А.М.* О создании машинного фонда тюркских языков // Советская тюркология. 1988. №2. С. 98-101.

5. *Асқар Құдайбергенұлы Жұбанов.* Қазақ сөзінің **«Тіл – қазына»** деректер базасы және оның теориялық негіздері. База данных **«Тіл – қазына»** казахского слова и ее теоретические основы. – Алматы: «Арыс» баспасы, 2009. – 304 бет.

6. *Ершов А.П.* К методологии построения диалоговых систем: феномен деловой прозы. Препр. №156, ВЦ СО АН СССР. Новосибирск, 1979.

ҚАЗАҚ ӘДЕБИ ТІЛІНІҢ ЭЛЕКТРОНДЫҚ КОРПУСЫН ТҮЗУДІҢ ТЕОРИЯЛЫҚ БАСТАУЛАРЫ

ҚР БҒМ ҒК А.Байтұрсынұлы атындағы Тіл білімі институты

атты профессор Ж.А.Манкееваның 60 жылдығына арналған
республикалық ғылыми-теориялық конференция.
Алматы-2010. – 191-197-бб.

Ең алдымен «тіл корпусы» ұғымының анықтамасына тоқталайық. Корпус – бұл электрондық пішіндегі мәтіндер жиынтығы. Корпус бойынша зерттеуші өзіне қажетті сөздерді, сөз тіркестерін, грамматикалық формаларын арнайы құрастырылған іздестіру жүйесі арқылы іздеп тауып, олардың мағыналарын аша алады. Ондай корпустардың бірнеше түрлерін кездестіруге болады, мысалы, бір ғана автордың корпусы, бір ғана кітаптың корпусы, бір ғана стильдің корпусы және т.б. Ондай корпустардың ең күрделісі және көлемдісі – белгілі бір тілдің Ұлттық корпусы. Ұлттық корпус сол тілдің өмір сүрген белгілі кезеңіндегі сан алуан жанрын, стилін, аймақтық, әлеуметтік нұсқасын және т.б. да түрлерін қамтиды. Аталған корпус түріне интернет желісіндегі 2003 жылдан бері өзіне жүктелген қызметті ойдағыдай атқарып келе жатқан «Орыс тілінің Ұлттық корпусын» атауға болады. Қазіргі кезде орыс тілінің Ұлттық корпусының жалпы көлемі 140 млн. сөзқолданыстан тұратын әртүрлі мәтіндер бөлігін қамтиды. Болашақта мәтіндер корпусының көлемін 200 млн. сөзқолданысқа дейін көбейту жоспарлануда [1].

Орыс тілінің ұлттық корпусы басқа да тілдердің ұлттық корпустары сияқты мынадай екі маңызды ерекшеліктерге ие:

- біріншіден, корпусқа аса көлемді және әр кезеңдер бойынша шамалас көлемдегі мәтіндердің алынуы. Дәлірек айтсақ, біріншіден, орыс тілінің ұлттық корпусына барлық жазбаша және ауызша мәтіндер (көркем әдебиеттің әртүрлі жанрлары, көсемсөз жанры, оқу, ғылыми, іскағаздары, сөйлеу тілі, аймақтық тіл және т.б.) қамтылып, олар әр кезең бойынша сәйкес көлемде алынады;
- екіншіден, корпус құрамындағы мәтіндер айрықша сипаттағы қосымша ақпаратқа ие болады. Мұндай ақпарат шартты түрдегі белгілеулер арқылы көрініс табады (орысша аталуы – «разметка» немесе «аннотация»). «Разметка» терминін қазақшалап, «белгілеу» деп атауды жөн көрдік. Сонымен, белгілеу (разметка) корпустың негізгі сипаттамасы деуге болады. Осындай сипаттама арқылы корпус көптеген қарапайым түрдегі мәтіндер жиынтығынан (немесе мәтіндер кітапханасынан) айырым табады. Мәтіннің грамматикалық сипатына қатысты белгі түрлерінің көп болуы корпустың ғылыми және оқыту саласындағы құндылығын арттыра түседі. Мәселен, орыс тілінің ұлттық корпусында қазіргі кезде белгілеудің үш түрі кең қолданыс табады. Біріншісі – метамәтіндік, яғни мәтінді жалпы түрде қарастыруға қатысты сипаттаулар: мәтін авторы, мәтін жанры және т.б. ерекшеліктер; екіншісі мен үшіншісі – морфологиялық және семантикалық сипаттаулар: мәтінді жалпы түрде қарастырмай, оның әрбір сөзіне қатысты грамматикалық ерекшеліктерін шартты белгілер арқылы белгілеу (разметка).

Ұлттық корпус, ең бірінші кезекте, тілші-ғалымдарға сол тілдің лексикасы мен грамматикасын жан-жақты зерттеуге мүмкіндік тудырады. Ал

корпустың келесі міндеті тілдің ішкі салалық (лексика, грамматика, тіл тарихы және т.б.) аясына қатысты әртүрлі анықтағыштық рөл атқару.

Егер Ұлттық корпуста тілдік бірліктердің статистикалық сипаты да берілетін болса, ондай деректермен әдебиетшілер, тарихшылар және басқа да қоғамдық ғылымдардың сала мамандары пайдалана алады. Әрине, ұлттық корпустың қолдану аясы тілдерді ана тілі немесе шет тілі ретінде оқыту кезінде көбірек байқалады. Сондықтан қазіргі кезде көптеген оқулықтар мен оқу бағдарламалары мәтіндік корпустарға бағышталып құрастырылуда. Мәселен, мағынасы күңгірт сөздер мен грамматикалық формалардың қолдану ерекшеліктерін белгілі авторлардың шығармалары бойынша электрондық корпус көмегімен әрі тез, әрі ұтымды тексеруді шетелдік азамат та, оқушы да, оқытушы да, журналист те және жазушы да жүзеге асыра алады.

Жоғарыда сөз болған орыс тілінің Ұлттық корпусының даму барысын сөз етсек, ол ең алдымен XIX ғ. басынан XXI ғ. бастапқы кезеңін қамтиды деуге болады. Бұл кезең орыс тілінің әртүрлі әлеуметтік-лингвистикалық тұстарын бейнелейді – әдеби тіл, сөйлеу тілі, тұрмыстық тіл, ішінара диалектілік тіл. Корпустық қорға көркем әдебиеттің (проза мен драматургия, поэзия) мәдени маңыздылығы мол және тілдік тұрғыда тілші-ғалымдардың қызығушылығын тудыратын түпнұсқа түріндегі шығармалары енгізіледі. Бірақ ұлттық корпус тек көркем әдебиетке қатысты мәтіндерден ғана тұрмайды, ол сонымен бірге мәтін үлгілерінің басқа да жазба нұсқаларын (қазіргі кезеңде – ауызша нұсқасын да) қамтиды. Оған жататындар: мемуарлар, эсселер, көсемсөз стильдері, ғылыми-көпшілікке арналған және ғылыми әдебиеттер, жұрт алдында сөйлеген сөздер, жеке адамдар арасындағы хат алысу, күнделіктер, құжаттар және т.б. мәтіндер болуы мүмкін [1].

Корпустық лингвистика – қолданбалы лингвистиканың заман сұранысынан туындаған салаларының бірі. Сондықтан корпустық лингвистиканың қарқынды дамуын сапалы грамматикалық өңдеуден өткен және тұлғалы тілдік деректерге жеңіл қол жеткізуді қамтамасыз ететін ресурстарды құрастыру қажеттілігімен түсіндіруге болады. Осындай қажеттіліктің ең негізгісі деп қолданбалы лингвистика мәселелерінің ішінен бір тілден екінші тілге компьютер арқылы аудару жүйесін тілдік материалдармен қамтамасыз ету әрекетін атауға болады. Корпустық лингвистикада қолданыс табатын компьютерлік бағдарламалар әртүрлі табиғи тілдерді өңдеуге бағытталған. Қазіргі кезде түрлі тілдердің мәтіндер корпусын құрастыру әрекетінде қолданыс табатын мұндай бағдарламалар (NLP – natural language processina) барынша жетілдірілуде [2, 60].

Мәтіндер корпустары алғашында табиғи тілді өңдеу үшін жазылатын бағдарламалармен және оларды тестілеу кезіндегі өңделмеген (шала) күйінде көрініс табады. Бұл жердегі мәтіндер ұғымына ауызша сөйлеу тілі де және табиғи коммуникациядағы тіл де қамтылады. Қорыта айтқанда, барлық бүгінгі күнге белгілі корпустар мынадай төрт нұсқада жүзеге асады:

1) коммуникацияның әртүрлі аясын қамтитын мәтіндер (монолингвальдық корпус) – ұлттық корпус;

2) жоғарыда аталған корпусқа ұқсас құрастырылатын (репрезентативтігі сәйкес және ортақ корпус-менеджерімен) бірнеше ұлттық корпустарды қамтитын – салыстырмалы (немесе контрастивтік) корпус;

3) бір тілдегі мәтіндерді және оның басқа тілдегі (тілдердегі) аударма мәтінін қамтитын – параллель корпус немесе параллель мәтіндердің корпусы;

4) ұлттық корпустың құрамына еніп қана қоймай, өзінше жеке тұрып та корпус құрайтын – сөйлеу тілінің корпусы.

Сөйлеу тілі корпустары (кең мағынада – ауызша) ауызша тілдің тек аудиожазбасын (мысалы, қайсыбір тілдің диалектілер корпусы) немесе, керісінше, оның тек транскрипциялық не орфографиялық түрін қамтуы мүмкін (мысалы, Британдық ұлттық корпус массивінің бір бөлігі). Үйлесімді нұсқа ретінде корпуста аудиожазба мен оның стенограммасының (транскрипцияланған немесе орфографиялық түрде) қатар орналасқаны дұрыс деп саналады [2, 60-61].

Аталған нұсқалардың әрбіреуі өзіне тән ерекшелігі бар мәселелерді ғана шеше алады. Мысалы, салыстырмалы корпус контрастивтік аспектідегі тілдерді үйренуге мүмкіндік береді, сөйте тұра, параллель корпустар тәріздес, қазіргі машиналық аударма жүйелерінде деректер базасы (аударма мысалдарының базасы) ретінде қолданылады. Ал контрастивтік зерттеу кезінде салыстырмалы корпусты қолдану мақсатқа сай келе бермейді. Себебі екітілді параллель мәтіндерінде аударма жасауда сөзсіз орын алатын тіларалық интерференция (екітілдік жүйенің бір-біріне тигізер әсері) көрініс тауып, салыстыру нәтижесіне айтарлықтай дәрежеде кері әсер етуі мүмкін.

В.П.Захаров өзінің оқу құралында корпустардың толық түрдегі топтастыруын ұсынады [3, 13]. Төменде осындай топтастыру *1-кесте* түрінде көрініс тапты.

Кесте мәліметтеріне бойынша, корпустар, негізінен, кез келген коммуникация аспектісін модельдеуге жағдай жасайтынына көз жеткізуге болады. Корпусқа негізделген модельдердің ең басты құндылығы – корпуста орын алатын барынша көптілдік материалдар негізінде мүмкін болатын анықтық (шындық) пен заңдылыққа қол жетуі.

Кез келген корпус аннотацияланған корпус (шартты түрдегі грамматикалық белгілерді енгізу) болғандықтан, ол мәтіннің лингвистикалық сипатының көрінісін айқындай алады. Әдетте ондай сипаттама корпус құрамындағы бірліктерге қатысты мүмкін болатын тілдік талдау үлгісін қамтиды. Аннотацияланған корпус, оны пайдаланудың қарапайымдылығы мен көпфункционалдық мүмкіндікке ие болуымен де аса құнды. Көпшілік жағдайда, корпустарды аннотациялау әрекетінің автоматты және «қолмен» жүргізу түрлері қатар жүргізіледі. Ал автоматтанған аннотациялау әрекетін қанағаттандырарлық деп санау үшін қателік шегі үш пайыздан аспау керектігін ескеру қажет [2, 63].

Кесте 1. В.П. Захаровтың корпустар топтамасы

Белгісі	Корпустар түрлері
Деректер түрлері	Жазбаша Ауызша (сөйлеу тілі) Аралас
Мәтіндер тілі	Орысша, ағылшынша, түрікше, қазақша және т.б.
«Параллельдік»	Біртілді Екітілді Көптілді
Әдеби тілге қатыстылығы	Әдеби Диалектілік Ауызекі (разговорные) Терминологиялық Аралас
Жанр	Әдеби Фольклорлық Драмалық Публицистикалық
Жетімділік	Кедергісіз Коммерциялық Жабық
Міндеті	Зерттеулік Иллюстрациялық
Динамикалығы	Динамикалық (монитарлық) Статикалық
Белгіленуі	Белгіленген Белгіленбеген
Белгілену сипаты	Морфологиялық Синтаксистік Семантикалық Просодикалық және т.б.
Мәтін көлемі	Толықмәтіндік Фрагменттімәтіндік (үзіндімәтіндік)
Хронологиялық аспект	Синхронологиялық Диахронологиялық
Жалпылық	Жалпы Жеке (бір ғана жазушы немесе т.б. мәтін түрлері)
Құрылымы	Орталық және архивтік Ядеролық және перифериялық

Қазіргі кезде кейбір тілдерге қатысты жүзеге асырылған компьютерлік бағдарламалар қанағаттандырарлықтай дәрежеде жұмыс істейді. Мәселен, сөзформаларды лемматизациялау (лексема түріне келтіру) және оның қай сөз табына қатыстылығын анықтау әрекеттері ағылшын, француз және испан тілдері бойынша толығымен автоматтандырылған [4, 455]. Бірақ автоматтанған белгілеу нәтижелерін «қолмен» (автоматтандырылмаған) тексеру жұмыстары әлі де болса орын алып келеді. Мысалы, Британдық ұлттық корпустың негізгі массиві осылайша, яғни әрі автоматты, әрі қолмен тексеру арқылы аннотацияланған (белгіленген) болатын. Әзірше, көптеген тілдер бойынша, тілдік тұрғыда күрделі болып саналатын мәтіндерге шартты белгілерді енгізу жұмысы толығымен «қолмен» (автоматсыз) атқарылуда.

Мысалы, анафоралық және катафоралық референцияны құрастыруда қолмен белгі қою орын алды. Сонымен, мәтіндер корпустарына автоматты түрде белгі қою жұмысы әлі де болса өзін ақтай алмай келеді, яғни қате кету мүмкіндігі мол [2, 65].

Кесте 2. **Корпустарды белгілеу түрлері**

Аннотация (белгілеудің) түрлері	**Анықтама**	**Сілтеме**
1	**2**	**3**
Лемматизация	Сөзформаны лексемаға келтіру, лексеманы анықтау	**Кең тараған процедура.** Оның сипатталуын орыс тілінің Ұлттық корпусы сайтынан (www.ruscorpora.ru) және орыс әдеби тілінің корпусы сайты бойынша (*www.narusco.ru*) танысуға болады.
Сөз табына қатыстылығы (тэггер)	**Морфологиялық белгілеу:** грамматикалық формаларды анықтау, грамматикалық мағыналарды кодтау (шартты белгілер қою)	**Кең тараған процедура.** Оның сипат-талуын орыс тілінің Ұлттық корпусы сайтынан (www.ruscorpora.ru) және орыс әдеби тілінің корпусы сайты бойынша (www.narusco.ru) танысуға болады.
Парсинг	**Морфосинтаксистік және синтаксистік белгілеу:** тәуелділік таралымын (деревье зависимости) құру, синтаксистік конструкцияларын анықтау	**Кең тараған процедура.** Оның сипатталуын орыс тілінің Ұлттық корпусы сайтынан (www.ruscorpora.ru) және орыс әдеби тілінің корпусы сайты бойынша (www.narusco.ru) танысуға болады.
Семантикалық белгілеу	Семантикалық категорияларды және олардың лексикалық пен грамматикалық көрсеткіштерін бейнелеу тәсілдерін анықтау	**Әзірше кең таралмаған процедура.** Мұны Сингапурлық Корпустан бойынша танысуға болады [*Ng H.T., Lee H.B.* Integrating multiple knowledge sources to disambiguate world sense: an example-based approach. Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (Santa Cruz, Calif.) 1996. P. 40-47]
Дискурстық белгілеу	Дискурстық параметрлерді анықтау (дискурс түрлері және т.б.), екпін, үзіліс, түсіндірме (огаворка) түрлері арқылы просодика-лық белгілеулер қамтылады	**Жедел дамып келе жатқан сала** [*Botley S., McEnery A.* (eds.) Discourse Anaphora and Resolution. Studies in Corpus Linguistics. Amsterdam: John Benjamins. 2000]
Прагматикалық белгілеу	Прагматикалық параметр-лерді анықтау (сөйлеу акті-лерінің түрлері), анафора-лық белгілеулер түрлері қамтылады (есімдіктер референциясын анықтау)	**Жедел дамып келе жатқан сала** [*Short M.,, Culpeper J., Semino E.* Using a corpus for srylistics research: speech presentation // *Short M and Tomas J.* (eds). Using Corpora for Language Research. London: Longman, 1996]
Ақпаратты шығарып алу (IE – orlented annotation)	Корпус мәтінінің мазмұнына талдау жүргізу негізінде ақпаратарды іздеу және оны компьютер жадынан шығарып алу	[*Aone C., Bennett S.W.* Discourse tagging and discourse tagged multilingual corpora. Proceeding of International Workshop on Sharable Natural Language Resources (Nara), 1994. P. 71-77]

Ғалым Энтони МакЭнери өз зерттеуінде [4, 458] шетелдік дереккөздерге сүйене отырып, әртүрлі белгі қою мүмкіндіктерін ***2-кестеде*** көрсетілгендей сипаттайды. Осы сияқты корпустарға берілген аннотацияларды ғаламтордағы (интернеттегі) Ұлыбританияның Ланкастер университетінің мына сайтынан табуға болады [2, 65]: *http://www.Ling.lancs.ac.uk/monkey/ihe/linguistics/corpus2/2fral.htm.*

Орыс тіл білімінде морфологиялық белгілеу үш кезең бойынша іске асырылады: анализ-талдау (парсинг); сұрыптау; омонимдерді ажырату. Мұндағы бірінші кезеңде жасалатын лингвистикалық (морфологиялық) белгілеу әдісі MYSTEM программасы бойынша жүзеге асады. Бұл программа А.А.Зализняктың «Орыс тілінің грамматикалық сөздігіндегі» мәліметтерге негізделген және мұнда әрбір сөзформаның морфологиялық белгілерінің мүмкін нұсқалары ирек жақша ішіне жазылады.

Бірінші кезеңде сөзформаның мүмкін варианттары автоматты түрде түгелдей қамтылса, екінші кезеңде омоним сөзформалар анықталып, олардың тұсына сұрақ белгілері қойылып отырады. Бұл да автоматты түрде GRAMBAT программасына салынады. Үшінші кезеңде анықталған омонимдер, яғни бірдей сөзформалардың (ол сөз не сөздің формасы, яғни грамматикалық формалар болуы мүмкін) лексикалық не грамматикалық мағыналары мәтіндерден іздестіріліп, қай сөз табы екендігі автоматты түрде емес, қол жұмысы арқылы жүзеге асырылады [7, 42-46].

Қазіргі қазақ тіл білімінің даму үдерісі заман өзгерісіне сай, жаңа технологияны – компьютер мүмкіндіктерін ұлттық тілді зерттеу барысында кең түрде қолдануды қажет етуде. Осыған байланысты А.Байтұрсынұлы атындағы Тіл білімі институтында «Компьютерлік лингвистика» мен «Корпустық лингвистика» салаларына тікелей қатысты бар «Мәдени құндылықтар ретіндегі қазақ тіліндегі мәтіндер корпусы және сөздіктердің «Тіл – қазына» атты ұлттық компьютерлік қоры» атты тақырып бойынша ғылыми-зерттеу жұмысы 2009 жылдан бастап жүргізіле бастады. Ал бұл зерттеу жұмысының негізгі мақсаты – қазақ тілінің мәдени құндылығы болып саналатын толық мәтіндер корпустарының мәтіндеріне грамматикалық белгі-кодтар енгізіп, оның дербес түрдегі «Тіл – қазына» атты компьютерлік базасын құру. Ал аталған компьютерлік база коммуникация қажеттігін толығымен қанағаттандыра алатындай, үлкен көлемді мәтіндер мен сөздіктерге қатысты ақпараттық-сұраныс жүйесі түрінде көрініс табуы қажет. Қазақ тілінің көркем әдебиет мәтіндерінің компьютерлік корпусының нысандары ретінде Абай Құнанбаевтың, М.Әуезовтің, Ә.Кекілбаевтың, М.Мақатаевтың, М.Мағауинның толық шығармаларынан таңдама мәтіндер алу жоспарланды.

Қазақ мәтініне морфологиялық деңгейдегі белгі-код қоюға қатысты зерттеу жұмыстарын А.Байтұрсынұлы атындағы Тіл білімі институтының «Компьютерлік лингвистика» бөлімінің ғалымдары (С.Құлманов, О.Жұбаева, А.Жаңабекова және т.б.) жүзеге асыруда [5, 6, 7].

Қазақ тілінің көркем әдебиет стилі мәтініне қатысты белгілеу (разметка) жұмысынан келтірілген үзінді-мысал төмендегідей көрініс тапты:

Құнанбайдың алдында мысықтай үндемей отырып, былай шыға бере, бала-шаға, малшы-жалшыны сықпыртып боқтап жүру оның дағдысы болған.

Құнанбайдың {**Құнанбай** – зат: жалқы, күрделі, біріккен, деректі + Ø жекеше + септік, ілік}

алдында {**алд** – көмекші есім: дара, түбір + Ø жекеше + тәуелдік, 3-жақ + септік, жатыс}

мысықтай {**мысықтай** – сын: дара, туынды, қатыстық}

үндемей отырып {**үндемей отыр** – етістік: күрделі, тіркес // ***үнде*** – етістік: туынды, негізгі, салт + болымсыздық + көсемше / ***отыр*** – етістік: түбір, көмекші, қалып + көсемше}

былай {**былай** – үстеу: дара, негізгі, мекен}

шыға бере {**шыға бер** – етістік: күрделі, тіркес // ***шық*** – етістік: түбір, негізгі, салт + көсемше / ***бер*** – етістік: түбір, көмекші + көсемше}

бала-шаға {**бала-шаға** – зат: жалпы, күрделі, қос, деректі + Ø жекеше + Ø септік, табыс}

малшы-жалшыны {**малшы-жалшы** – зат: жалпы, күрделі, қос, деректі + Ø жекеше + септік, табыс}

сықпыртып {**сықпырт** – етістік: дара, түбір, негізгі, сабақты + көсемше}

боқтап жүру {**боқтап жүр** – етістік: күрделі, тіркес // ***боқта*** – етістік: туынды, негізгі, сабақты + көсемше / ***жүр*** – етістік: түбір, көмекші, қалып + тұйық етістік}

оның {**ол** – есімдік: дара, түбір, жіктеу + Ø жекеше + септік, ілік}

дағдысы {**дағды** – зат: жалпы, дара, түбір, дерексіз + Ø жекеше + тәуелдік, 3-жақ + Ø септік, атау}

болған {**бол** – етістік: дара, түбір, көмекші + есімше + Ø рай, ашық + Ø жіктік, 3-жақ, Ø жекеше}.

Ең алдымен айтпағымыз біз корпусқа енгізілетін мәтіндерге грамматикалық (морфологиялық) талдау жасауды мақсат етіп отырмыз. Берілген сөйлемдегі сөздерге жасалған осындай талдаудың негізгі принциптеріне, қойылған шартты белгілерге түсініктеме бере кетсек.

1) Мәтіндегі сөздер (қара жазумен) жеке бөлек алынады: **Құнанбайдың;**

2) Одан кейін ирек жақша қойылып, осы сөздің (қара жазумен) реестрге енетін тұлғасынан кейін «сызықша белгісі» арқылы негізгі грамматикалық сипаттамасы беріледі: **Құнанбай** – зат: жалқы, күрделі, біріккен, деректі;

3) + белгісі қойылып, осы сөзге жалғанған (жалғануы мүмкін) грамматикалық формасы (бірнешеу болса, араларына + белгісі қойылады) беріледі: + Ø жекеше + септік, ілік. Мұнда Ø символы «нөлдік форма» дегенді білдіреді, яғни бұл жерге көптік формасы жалғануы мүмкін (Құнанбайлардың);

4) сөз құрамындағы барлық грамматикалық формалар берілгеннен кейін ирек жақша жабылады.

Морфологиялық талдау, негізінен, контекске байланыссыз жасалатыны белгілі. Алайда мұндай талдау барысында қазақ тілінің құрылымдық (әсіресе синтаксистік) ерекшеліктерін ескермеске болмайды. Мысалы, келтірілген сөйлемдегі *шыға бере* тіркесінің құрамындағы сөздерді жеке алып талдауға болмайды. Өйткені *беру* етістігі бұл жерде көмекшілік қызметте қолданылғандықтан, оның мәтіндегі грамматикалық қызметін жеке алып талдағанда ашылмай қалады. Осы себепті біз бұл тіркесті бірге алып төмендегіше беріп отырмыз: **шыға бере {шыға бер** – етістік: күрделі, тіркес // ***шық*** – етістік: түбір, негізгі, салт + көсемше / ***бер*** – етістік: түбір, көмекші + көсемше}. Мысалдың талдану үлгісі:

1) Мәтіндегі сөз тіркесі (қара жазумен) бірге алынады: **шыға бере;**

2) Одан кейін ирек жақша қойылып, осы сөз тіркесінің ең кіші тіркесу үлгісі (қара жазумен) алынып, сызықша арқылы оның негізгі грамматикалық сипаттамасы беріледі: **шыға бер** – етістік: күрделі, тіркес;

3) // белгісі қойылып, осы сөз тіркесінің құрамындағы бірінші сөзден (қара көлбеу жазумен) кейін сызықша арқылы оған жалғанған (жалғануы мүмкін) грамматикалық формасы (бірнешеу болса, араларына + белгісі қойылады) беріледі: // ***шық*** – етістік: түбір, негізгі, салт + көсемше;

4) / белгісі қойылып, осы сөз тіркесінің құрамындағы келесі сөзден (қара көлбеу жазумен) кейін сызықша арқылы оған жалғанған (жалғануы мүмкін) грамматикалық формасы (бірнешеу болса, араларына + белгісі қойылады) беріледі: / ***бер*** – етістік: түбір, көмекші + көсемше. Егер алынған сөз тіркесі екі сөзден асса, сөздерді талдауда осы соңғы үлгі қолданылады.

5) сөз тіркесінің құрамындағы соңғы сөздің құрамындағы барлық грамматикалық формалар берілгеннен кейін ирек жақша жабылады.

Біз бұл мақаламызда бір ғана сөйлемге жасалған морфологиялық талдау үлгісіне қысқаша тоқталдық. Қазақ тілінің грамматикалық сипатына байланысты осындай талдау үлгілерінің ерекшеліктері туралы жеке мақала түрінде баяндау қажет деп ойлаймыз.

Қазақ тілінің лексикалық қорының үнемі артып отыру үдерісін лексикографиялық ізденістер мен баспадан шығып жатқан біртілді, екітілді сөздіктер түрлерінен де байқауға болады. Осыған дәлел ретінде, ең алғаш жарық көрген екі томдық, он томдық және қазіргі кезде құрастырылып жатқан он бес томдық «Қазақ әдеби тілі сөздіктерінің» лексикалық көлемдерінің айтарлықтай ұлғаю сипатын айтуға болар еді.

Сонымен, жоғарыда сөз болған жайттар қазақ тілін зерттеуде корпустық лингвистика саласына көп мән берілуін, оның әлемдік дәрежедегі теориялық және практикалық жақтарын жан-жақты зерделеу қажеттігін көрсетеді.

Әдебиет

1. Национальный корпус русского языка // http://www.ruscorpora.ru.

2. *Овчинникова И.Г., Угланова И.А.* Компьютерное моделирование вербальной коммуникации : Учеб.-метод. пособие. –М.: Флинта: Наука, 2009. – 136 с.

3. *Захаров В.П.* Корпусная лингвистика. –СПб., 2005.

4. *Mc Enery A.* Corpus Linguistics // Mitkov R. (ed.) The Oxford Handbook of Computational Linguistics. –Oxford: Oxford University Press, 2004. –P.448-463.

5. *Жұбанов А.Қ.* Корпустық лингвистика – қазақ тіл білімінің жаңа бағыты // Тілтаным (Языкознание). –Алматы, 2009, №2 (34). –3-10 бб.

6. *Жұбаева О.* Корпустық лингвистиканың теориялық және практикалық құндылығы / Қазақ тіл білімінің өзекті мәселелері // Профессор Құдайберген Жұбановтың 110 жылдығына арналған халықаралық ғылыми-теориялық конференция материалдары. –Алматы, 2009. –251-255 бб.

7. *Жаңабекова А.Ә.* Тілдік корпус ұғымының мәні, сипаттамасы // Тілтаным (Языкознание). –Алматы, 2009, №4 (36). –42-46 бб.

8. http: //www.Ling.lancs.ac.uk/monkey/ihe/linguistics/corpus2/2fral.htm;

9. http: //www.narusco.ru.

МӘТІНДІ АВТОМАТТЫ ӨҢДЕУ ЖҮЙЕСІНІҢ ҚҰРЫЛЫМЫ ЖАЙЫНДА

Профессор ҚазКСР ҒА академигі Нығмет Тінәліұлы Сауранбаевтың туғанына 100 жылдығына арналған «Түркі тілдерінің грамматикасы: қазіргі ахуалы мен болашағы» атты халықаралық ғылыми-теориялық конференция материалдары. –Алматы: А.Байтұрсынұлы атындағы Тіл білімі институты, 2010. – 159-166-бб.

Ғылым мен техниканың даму деңгейі көбіне ғылыми-техникалық ақпараттардың қаншалықты дер кезінде және сапалы түрде өңдеу дәрежесімен анықталады. Осыған қатысты мәтінді өңдеу тұсында қажетті ақпаратты іздестіру мәселесі туындайды. Сонымен қатар ЭЕМ-ді (компьютерді) қолдану ісі «жасанды интеллект» саласына қатысты мәселелерді шешуді де қажет етеді.

Бұл аталған екі мәселе де мәтіндік ақпаратты талдауға тікелей байланысты болуымен бірге табиғи тілдегі мәтіндерді өңдейтін автоматтанған жүйені қолдануды талап етеді.

Мәтіндерді автоматты өңдеу (МАӨ) ана тіліміздегі және шет тіліндегі ақпаратты сараптауды, реферат алуды, аударма жасауды жүзеге асыратын күрделі іс-әрекеттерге негіз болады. Сол сияқты диалогтық сұрақ-жауап түріндегі оқыту жүйелері де МАӨ-сіз әрекет ете алмайды.

Аталған мәселелердің жиынтығын жүзеге асыру үшін мәтіндерді автоматты өңдеудің әртүрлі деңгейлерімен әрекет ететін лингвистикалық автоматтарды (ЛА) қолданысқа енгізу керек болады. ЛА біздің

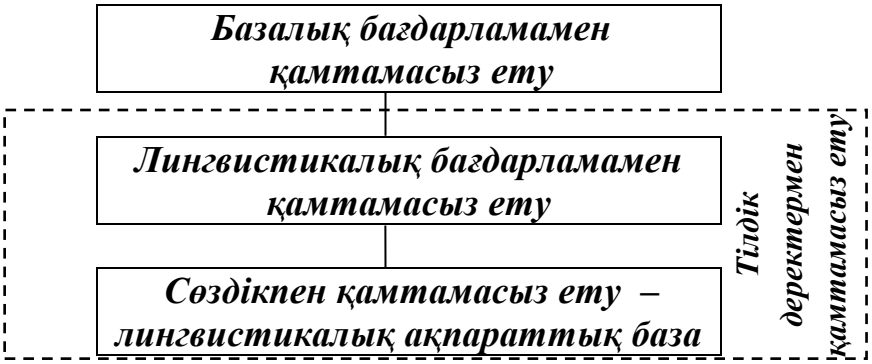
қалағанымыздай болуы үшін адамдардың лингвистикалық іс-әрекетін толығымен қайталай алатындай немесе модельдей алатындай болуы қажет [1]. ЛА-ны мәтіндерді автоматты өңдеу жүйесін ұзақ мезгіл жетілдіру нәтижесінің шегі деп те қарастыруға болады.

ЛА-ны құрастыру ісі лингвистиканың, кибернетиканың және есептеуіш техниканың даму деңгейлерімен тікелей байланысты болуымен бірге, ол психология, нейролингвистика және семиотика ғылымы салаларының жетістіктерімен де қаруланады [2; 3; 4; 5].

Адам миында туындайтын тілдік бірліктер арасындағы байланыс құрылымы мен ақпараттың қалай сақталатындығын меңгермейінше, тілші маманның іс-әрекетін модельдеу мүмкін емес. Сонымен бірге ЛА-ның жүзеге асырылған моделін құру әр уақытта саналы түрде мәтінді қабылдау мен оны туындатуды танып-білу түсінігіне негізделеді [6].

Әрине, мұндай баламаны білу ЛА-ны құруда, оны қалай болса солай көшіріп алу деген сөз емес. ЛА-ның құрылымы бөлек болуы да мүмкін, бірақ оның модельдеу қызметі балама әрекетімен барабар болуы қажет. Мысал ретінде, ұшақтың моделі мен оның табиғи нысаны – құсты айтуға болар еді.

Қазіргі кезде біз қазақ тілінің лингвистикалық автоматын құрастырып шығу ісінде әлі де болса білім қорын жинақтап бола қойғанымыз жоқ. Бірақ бұдан ЛА-ны құрастыруды жақсы жағдай туындағанша тоқтата тұру қажет деген ұғым тумауы тиіс. Қазақ тіліне қатысты лингвистикалық автоматтар бүгіннің өзінде-ақ қажет. Бірақ тәжірибелік модельдер – «ойыншықтар» емес, өндірістік коммерциялық жүйелерге қажет. ЛА-ны құрастыру деңгейлік, модельдік және ашық жүйе түрінде біртіндеп, итеративті әдіспен (қайталанатын әрекет әдісі) жүргізілгені жөн болады.



1-сурет. Лингвистикалық автоматтың (ЛА) жалпы құрылымы

Бұл жүйе негізгі үш құраушыдан тұрады. Ең әмбебаптық сипаттағы құраушы – базалық бағдарламамен қамтамасыз ету компоненті болып саналады. Ол лингвистикалық автоматты тілдік деректермен қамтамасыз етуді қалыптастыратын келесі екі компонентті құру мен іске қосу қызметін атқарады. ЛА-ның лингвистикалық деректермен қамтамасыз етуін қалыптастыру бағдарламалық кешен мен ақпараттық база негізінде жүзеге асады. Бағдарламалық кешен деп отырғанымыз, мәтіндерді автоматты түрде өңдейтін бағдарламалар жиынтығы немесе лингвистикалық

бағдарламалармен қамтамасыз ету әрекеті. Ал ақпараттық база дегеніміз – сөздіктер жиынтығы түрінде құрылған ЛА-ны сөздіктермен қамтамасыз ететін жүйе. Лингвистикалық автоматтың жалпы құрылымы ***1-суреттегідей*** көрініс табады.

ЛА-ның архитектурасында деңгейлік тәсіл мен оны жүзеге асыруға қажетті модульдік принциптің бір-бірімен тығыз байланыста болғанымен, оларды бөлек қарастыруға болмайды. Жұмысқа арналған мәтінді қабылдау мен туындату моделі негізінде анықталатын талдау деңгейлері ЛА-да қарастырылатын шаралардың, деңгейлердің және оларға қажетті ақпараттардың реттілігін айқындайды. Сондықтан ЛА-ны сөздік ақпаратымен қамтамасыз ету модульдік пен ашықтық және деңгейлік тәсіл принциптерінде құрылуы қажет.

Сөздікпен қамтамасыз ету әртүрлі мәтінді автоматты өңдеу (МАӨ) жүйелері жиынтығының базасы (қоры) болып табылады. Ал онсыз табиғи тілдегі (ТТ) ақпаратты өңдеу мүмкін де емес [7].

Лингвистикалық автоматты лингвистикалық бағдарламамен қамтамасыз етуге қатысты деңгейлердің иерархиясын анықтағанда «жоғарыдан төмен бағытталған» принцип базалық болып саналады. Осының негізінде толық мәтінді талдаудан, оның үзінділерін бөліп алып талдауға біртіндеп (реттілікпен) көшу іске асырылады, ал содан кейін барып үзінділер ішіндегі әр сөйлемді талдау, мысалы, оларды аудару немесе автоматты өңдеудің басқа түрлері жүзеге асады [8; 9; 10]. МАӨ-нің нәтижесі компьютердің мәтінді «түсіну» деңгейі арқылы анықталады.

Адамның да, лингвистикалық автоматтың да мәтін мазмұнын дұрыс ұғынуға негіз болатын қоршаған орта жайындағы алдын ала қалыптасатын түсінік (жағдаят). Ондай түсінікке ішкі байланыста тұратын мәтін ақпаратын талдау арқылы қол жетеді. Жағдаятты құраушылар арасынан бас бөліктерін бөліп алып, олардың дифференциалдық белгілерін анықтау және шартты түрде мәтін түсінігіне жатқызатын адам тәжірибесінің категориялар жүйесіне кірістіру – талдаудың нәтижесі болып табылады [11].

Жаңа ақпаратты қабылдағанда, оның жекелеген бөліктері мен араларындағы байланыстары реципиенттің (қабылдаушының) тезаурусындағы бұрыннан жинақталған әлеуметтік біліммен салыстыру әрекеті жүзеге асады деп болжам жасауға болады. Сонымен бірге оның нәтижесінде қабылдаушының тезаурусы қабылданған хабарламаның жаңа элементтерімен және олардың арасындағы байланыстармен толықтырылады.

Талдау жүргізілетін деңгей реттерінің (мәтін, үзінділер, сөйлемдер, сөз тіркестер, сөздер) бірінші кезеңінде ең төменгі лингвистикалық деңгей тірек етіп алынады. Әріптік және буындық графиканы пайдаланатын тілдерде (үндіеуропалық, түркі, финн-угорлық және т.б.) мұндай деңгей ретінде сөзформа алынады. Ал иероглифтік жазба тілдерінде (қытай, ішінара жапон) бұл деңгей – иероглифтермен таңбаланатын буынморфтар деңгейі.

Сондықтан егер болжам жасау мен ЛА-ның құрылымын зерттеу «жоғарыдан төмен бағытында», яғни ең жоғарғы деңгейден төменгі деңгейге қарай жүргізілсе, онда ЛА-ны құрастыру мен оны сөздікпен қамтамасыз ету

әрекетін базалық бірліктердің төменгі деңгейінен бастап жүргізуге тура келеді.

Сөздікпен қамтамасыз ету құрылымын жоспарлау базалық бірлікті жан-жақты сипаттауға негізделеді. Сондықтан әрбір базалық бірлікке тіркеліп жазылатын ақпарат талдаудың кез келген деңгейіне қатысты мәселелерді шешетіндей дәрежеде болуы қажет.

Құрылымдық әдіске сәйкес бөлініп алынатын нақты талдау деңгейін сипаттау осы деңгейді құрастыратын талдау қабаттарының сипаттауларына ажыратылады. Әрбір сипаттауға сол деңгейдегі немесе қабаттағы талдау алгоритмі және деректер жиыны сәйкес келеді. Олар ЛА-ны сөздікпен қамтамасыз ету кезінде ескерілуі қажет.

Сонымен, кез келген сөздікпен қамтамасыз етілген ЛА ақпараттық база болып табылады және ол иерархиялық, модульдік сәйкестікте және ашық жүйе түрінде жүзеге асуы керек.

Қазіргі кезде мәтінді автоматты өңдеу (МАӨ) жүйелерінің сан алуан түрлері бар. Олар бір-бірінен тек алға қойылған мәселелерімен, жүзеге асыру деңгейімен, пәндік аямен ғана айырым тауып қоймай, сонымен бірге ақпараттық базамен, сөздіктермен қамтамасыз етілу негіздерімен де ерекшеленеді. Бұл, ең алдымен, зерттеушілер ұжымының әркелкілігімен түсіндірілсе, екіншіден, МАӨ жүйелерін лингвистикалық қамтамасыз етудің теориясының әлі де болса ойдағыдай болмауымен түсіндіріледі.

Енді МАӨ-ні лингвистикалық қамтамасыз ету негіздерінің кейбір мәселелеріне тоқталайық. МАӨ-нің әртүрлі мәселелерін шешуді қажет ететін лингвистикалық қамтамасыз ету бір ғана негізде болып, базалық бірліктердің толық бейнеленуімен және құрылымдық пен жүзеге асыру бағдарламасымен сипатталуы қажет [12]. Сондықтан барынша көп әмбебаптық сипаттағы лингвистикалық ақпараттық база (ЛАБ) жағдайын зерттеп, бірден-бір жүйені іске қосу аса маңызды. Мәтінді автоматты өңдеудің лингвистикалық ақпаратты молынан қажет ететін ең күрделі мәселелерінің бірі – машиналық аударма (МА).

Мәтінді өңдеу жүйелерінде деректер базасы (ДБ) лингвистикалық ақпараттық база түрінде құрылады. ЛАБ лексикалық бірліктердің не бір ғана автоматты сөздігі (АС) мен синаксистік (семантика-синтаксистік) құрылым түрінде, не ақпараттың әртүрлі типтеріне ие болатын сөздіктер жүйесі түрінде жүзеге асуы мүмкін.

ЛАБ-қа қойылатын негізгі талаптар (әсіресе егер ол АС жүйесі түрінде көрініс тапса) мынаған саяды:

- ЛАБ ғылыми-техникалық ақпаратты өңдеудің (библиографиялық іздеу, аннотациялау, рефераттау мен құжаттарды компьютер арқылы аудару) автоматтанған жүйелерін тілдік тұрғыда қамтамасыз етуге байланысты мәселелерді шешуге бағытталған әмбебаптық сипатта құрылып, адам-компьютер жүйесіндегі жасанды интеллектіні құрастыру бағытында жасалуы қажет;

- ЛАБ әртүрлі автаматтанған ақпаратты өңдеу жүйелерінің бірден-бір негізі болып қалуы қажет;

- ЛАБ әртүрлі мәтінді автоматты өңдеу (МАӨ) жүйелерінің алдына қойылатын мәселелерге қатысты оңтайлы (оптималды) ұйымдастырылуы керек;

- ЛАБ-тың ең маңызды сипаттамасы болып саналатын әмбебаптығы мына жағдаймен анықталады: мәтінді өңдеудің барлық түрлерінде (ықтималдық, құжаттың мағынасын детерминдік және тезаурустық тануда, машиналық аудармада, мәтінді ықшамдауда, табиғи тілдегі адам-компьютер аралық диалогте және т.б.) әдетте тек бір ғана лексика-логикалық және грамматикалық ақпарат қажет болады. Мұндай құрамда ұйымдасқан ақпараттық базаның өзін қайта құрастырмай-ақ, әртүрлі МАӨ жүйелерінде пайдалануға болады.

Егер ЛАБ әмбебаптық база ретінде жүзеге асырылатын болса, онда оны әртүрлі МАӨ жүйелерінде пайдалану дегеніміз – ЛАБ-ты ондай жүйелердің біртұтас кешендік құралы деп санау. Әрбір нақты МАӨ-нің мәселелері үшін ЛАБ-ты құрастыруға арналған зерттеулер аса күрделі және көп еңбек сіңіруді қажет ететін мәселе болғандықтан, біртұтас және көптеген мақсатқа арналған ЛАБ құрастыру ісі оған кететін шығынды азайтады.

Үйлесімді (оптималды) ұйымдасқан ЛАБ-тың құрылымында модульдік сипат алдын ала қарастырылады, яғни онда ЛАБ өзара қатаң байланыста емес, топтамалардың (блоктардың) жиынтығы ретінде жүзеге асады. Мұндай модульдік сипат ЛАБ-ты жеке модульдерінің дайын болуына қарай құрастыруға мүмкіндік туғызады, ақпараттардың қайталануына жол бермейді және МАӨ мәселелерін реттілікпен, кезең-кезеңімен шешуге мүмкіндік жасайды.

ЛАБ-тың барлық деңгейлерінде реттілікпен жүзеге асатын модульдік принцип іштей базалық модуль-топтамаларын бөліп алуды және олардың арасындағы иерархиялық байланысты анықтауды қажет етеді. Мұндай иерархия мәтінді автоматты өңдеу (МАӨ) жүйелерінің қызметіндегі жалпы принциптерімен анықталады. Олардың ең негізгілері «төмен қарай түсетін» (нисходящий) принцип болып табылады. Соңғы айтылғанның мәнін былайша түсінуге болады. Күрделі кибернетикалық жүйелерді құрастыру және олардың қызмет етуі деңгейлік иерархиялық құрылымның негізінде белгілі реттілікпен жоғарғы басқару топтамаларынан (блоктарынан) бастап төменгі ақпарат деңгейіне қарай бағытталып жүзеге асады. Нәтижесінде жоғарғы деңгейлер төменгілердің жұмысын басқарып отырады. Сонымен иерархияның жоғарғы деңгейінде басқару модулі құрастырылады. Бұл модуль «диспетчер» рөлін атқаруымен бірге бағдарламалық талдау модульдерін рет-ретімен іске қосады. Олардың әрбіреуі ЛАБ-тың сәйкес топтамаларынан алынатын ақпаратқа сүйене отырып қызмет атқарады.

Бірақ МАӨ жүйесін құрастыру мен үйлестіру кезінде екі модульдік құрылымның – бағдарламалық кешен (жиынтық) мен ЛАБ-тың арасында қарама-қайшылық (конфронтация) пайда болады. Себебі кезекті бағдарламалық модульге көшу әрекеті алдын ала ең төменгі сөздік модуліне қатынас жасауды қажет етеді. Жүйе жұмысының бірінші кезеңінде жоғарғы модуль ЛАБ-қа қатынас жасамай тұрып және талдауға түскен мәтіннің

бірлігін сипаттайтын қайсыбір ақпаратты одан шығарып алмай тұрып, өз қызметін бастай алмайды. Бұдан кейін МАӨ жүйесінің қызмет ету барысы жоғарғы модульдердің әрекеті нәтижесінде төменгі ақпараттық деңгейден рет-ретімен шығарылып алынатын ақпаратты өңдеуге негізделеді.

Жоғарыда айтылғандарға байланысты МАӨ-де жүзеге асатын әрекеттер (процедуралар) бағдарлама жазуға үйлесімді «төмен қарай бағытындағы» және оған баламалы «жоғары қарай бағытыдағы» әдістер бір-бірімен өзара «келісімге» келулеріне тура келеді. Шындығында, жүйенің жоғарғы деңгейлерін жете зерттеудің алдында ЛАБ-ты жетілдіру жұмысына қатысты зерттеулер жүргізілгені жөн, ал жүйе қызметі ЛАБ-тың ең төменгі топтамаларына (блоктарына) қатынас жасамай тұра алмайды және одан кейін барып жоғарғы басқару модульдеріне қайта оралуы қажет болады.

Іске қосылған ЛАБ-тың көпшілігінде ядролық модуль-топтама (блок-модуль) ретінде лексикалық бірліктердің автоматты сөздігі (АС) орын алады. АС-тар сөздік бірлігі (СБ) мен олардың сипаттауларынан тұратын сөздік мақалалардың жүйесі түрінде құрастырылады. Мұндай АС-тың құрылымын анықтау үшін сөздік бірлігін таңдау, лингвистикалық бірліктің сипаттауын құрайтын ақпарат жиынтығы, сөздік мақаланы ұйымдастыруға қажетті үйлесімді әдіс-тәсілдерге қатысты мәселелердің шешімі табылуы қажет. Ол үшін мына талаптар орындалғаны жөн болады: ЛАБ-қа енетін ақпарат, онда бұрыннан да бар мәліметтерді қайталамауы керек [7]; ЛАБ-тағы бірліктер ішкі тұтастықта болулары керек; мәтіндік сөз ЛАБ бірлігімен ең оңай жолмен сәйкестікке қойылуы керек.

Сөздік бірлігін таңдау нақты тілдердегі сөз бен оның формаларының туындауымен және мәтінді мағыналық бірліктермен өрнектеумен де тығыз байланысты. Сондықтан бұл мәселені шешуде екіжақты ұстаным қажет болады: мәтін ішінен мақсатқа сай қандай лингвистикалық бірлікті (ЛБ) бөліп алу және қандай ЛБ-ны автоматты сөздікте (АС) сақтаған жөн.

Сөздік бірлігін таңдау мәселесін шешу модельденетін тілдің құрылымдық ерекшелігіне қарай анықталады. Ол үшін, ең алдымен, алға қойған мақсатқа сай, мәтін ішінен бөліп алынатын және сөздік бойынша сәйкестендірілетін мәтін бірлігі мен оған сәйкес АС бірлігін белгілеп алу қажет болады. Себебі мәтінді ЭЕМ-де өңдеу кезінде мәтіннің ең кіші бірлігінің шегін анықтау үшін қарапайым және бірмәнді критерий (формалды белгі) қажет болады. Бір жағынан қарағанда, оған сөздер арасындағы «бос орын» (пробел) жеткілікті де сияқты, бірақ бос аралық бойынша бөлініп алынатын мәтіндік бірлік тұрпат меже (план выражения) ретінде әр уақытта мазмұн межемен (план содержания) сай келе бермейді, яғни мұнда асимметрия бірмәнді критерий болуға «сенімсіздік» тудырады. Үндіеуропа, түркі және финн-угор тілдерінде бұл «сенімсіздік» толық мағыналы идиомаларды, күрделі көмекші сөздерді, жалғаулықтарды және т.б. мәтіннен бөліп алуда байқалады. Ал иероглифті пайдаланатын тілдерде аталған критерийдің «сенімсіздік» дәрежесі ұлғая түседі.

Сонымен мәтіндік сөздерге «бос орын» шекара болу үшін бір мағынаға ие «аналитикалық» сөздерді (сөз тіркестер мен иероглифтік тіркестер)

мәтіннен бөліп алуға қатысты арнайы зерттеулер жүргізілуі қажет болады. Бұл жағдайда мәтіннің негізгі бірлігі ретінде толық пішіндік сипаттағы сөзформа алынады [13]. Мәтіндік лингвистикалық бірліктерді (ЛБ) автоматты сөздік (АС) бірліктерімен сәйкестікке қойғанда ЛБ-ны тану мен теңдестіруге сөзформа бірлігі негізге алынады.

Әдебиет

1. *Пиотровский Р.Г.* Инженерная лингвистика и теория языка. – Л., 1979.
2. *Иванов В.В.* Чет и нечет. Ассиметрия мозга и знаковых систем. – М., 1978.
3. *Лурия А.Р.* Высшие корковые функции человека. – М., 1969.
4. *Лурия А.Р.* Основные проблемы нейролингвистики. – М., 1975.
5. *Пиотровский Р.Г.* Лингвистические аспекты исскуственного разума// ВЯ, 1981, №3.
6. *Marinov V.* Attention Hypothesis. – In: Empirical Semantics. – Bochum, 1981.
7. *Cercone N., Mercer R.* Design of Lexicons in Some Natural Language Systems. – ALLC Journal, vol.1, 1981, №2.
8. *Бондарко А.В.* Грамматическое значение и смысл. – Л., 1978.
9. *Волошин Г.Я.* Вопросы теории иерархических распознающих систем// В кн.: Семиотика и информатика. Вып. II. – М., 1979.
10. *Зимняя И.А.* Смысловое восприятие речевого сообщения. – В кн.: Смысловое восприятие речевого сообщения. – М., 1976.
11. *Лурия А.Р.* Предисловие к кн. Дж.Брунера «Психология познания». – М., 1977.
12. *Cercone N., Goebel R.* Approaches to Knowledge Representation. – In: Empirical Semantics. – Bochum, 1981.
14. *Смирницкий А.И.* Содержание и задачи лексикологии// Общее языкознание. Хрестоматия. – Минск, 1976.

МӘТІНДІ АВТОМАТТЫ ӨҢДЕУ ЖҮЙЕСІНДЕГІ ЛИНГВИСТИКАЛЫҚ АҚПАРАТ БАЗАСЫ

ҚР БҒМ ҒК А.Байтұрсынұлы атындағы Тіл білімі институты. Аса көрнекті ғалым, филология ғылымдарының докторы, профессор, ҚазКСР ҒА корр.-мүшесі Ахмеди Ысқақұлы Ысқақовтың туғанына 100 жыл толуына арналған «Қазіргі қазақ грамматикасы мен терминологиясын зерттеулердің негізгі бағыттары» атты халықаралық ғылыми-теориялық конференция материалдары. –Алматы: 2010. – 51-55-бб.

Мәтінді автоматты өңдеу (МАӨ) жүйесімен әрекет ету кезіндегі алғашқы «кіріс сөздігінің» бірлігі ретінде аналитикалық және флективті-аналитикалық тілдер түрлері үшін сөзформаны қолдану ең ыңғайлы деп саналуда. Нәтижесінде синтаксистік талдауға дейінгі талдау деңгейінде морфологиялық талдау алгоритмін қатыстырмай-ақ, сөздік бойынша қарастырылатын сөзге тіркелетін барлық ақпаратпен (қосымшалармен) танысуға мүмкіндік туады [1]. Сондықтан сөзформалардың жасалу тәсілдерінің тұрақты немесе тұрақсыздығына қарамай, қосымша түрлері тікелей автоматты сөздікте (АС-та) сақталады және оларды мәтіндегі «егіздерімен» теңдестіру ең жоғары жылдамдықпен жүзеге асады. Бірақ флективті тілдермен әрекет еткенде компьютердің жылдамдықтан ұтатын тұстары құнсызданады. Себебі АС бірлігі ретінде сөзформаны пайдалану сөздіктің көлемін ұлғайтуға әкеп соғады. Мысалы, орыс тіліндегі зат есім сөздердің форма өзгертуі 12-ге дейін, ал етістік сөздерде 188-ге дейін жетеді. Ал түркі тілдерінде атаушы тілдік бірліктердің форма өзгерту саны, теориялық жағынан алғанда, шексізге дейін жетеді деуге болады.

Міне, осы айтылғандарға байланысты мұндай парадигмаларды «ықшамдау» (компрессиялау) қажеттігі туындайды. Бұл әрекет екі түрлі тәсілмен жүргізілуі мүмкін: лингвистикалық (алгоритмдік) және техникалық. Лингвистикалық тәсілде негізгі назар сөздіктің бірлігін таңдауға аударылады, яғни бірліктері қайсыбір инварианттар болатын автоматты сөздікке (АС-ға) көшу.

Синтетикалық (флективтік және агглютинативті) тілдер үшін компрессиялаудың ең тиімді тәсілі автоматты сөздікті (АС-ты) негіз сөздер ретінде ұйымдастыру. Себебі канондық формадағы сөздік пен морфемдік сөздіктер айтарлықтай қолданыс таба алмайды деуге болады. Оларды пайдалану көптеген негіз түрлерін және негізден формаларды туындатуды қажет етумен қиындық туғызуда [2]. Негіз сөздердің «кіріс сөздіктерінде», әдетте, бірліктің екі түрі ажыратылады: алмасу немесе супплетивтік әдіс бойынша туынды формалар құрайтын шын мәніндегі негіз сөз (собственные основы) және ерекше форма деп аталатын түрі [3; 4]. Аталған бірліктердің екі түрі де АС-қа негіздермен бірдей дәрежеде тіркеледі. Сонымен бірге кейбір тілдерге қатысты парадигма үшін ерекше форманы пайдаланғанның өзінде де тек негізді ғана енгізу жеткіліксіз болады. Аса дамыған флективті тілге жататын орыс тілі үшін де АС негіз сөздер сөздігіне сүйенеді. АС-тың құрылымы бірліктердің неше түрлі жиынтығынан тұрады. GETA тобының жүйесінде орыс тілінің «кіріс сөздігінің» ұйымдасуы жеті сөздікті өз құрамында сақтауды қарастырады: екі сөздік негіздерден тұрса, келесі екеуі – өзгеріске ұшырамайтын сөздер мен ерекше формалар сөздіктері, соңғы екеуі – қосымшалар мен сөз орамы сөздіктері.

Орыс тілінің автоматты сөздігі (АС) жүйесінің көбірек зерттелуіне байланысты әрі қарайғы баяндауымызда сол сөздіктер құрылымына көбірек назар аударамыз.

Сөздік бірлігін таңдаудың прагматикалық негізі жайлы айтатын болсақ, сөздік бірлігін таңдау нақты МАӨ түрімен анықталады. Ал ЛАБ осы МАӨ-ні

жүзеге асыру бағытында қызмет етеді. Машиналық аударма (МА) үшін құрастырылатын автоматты сөздіктің (АС-тың) құрамына аса көп лингвистикалық бірліктердің (ЛБ-ның) енетінін ескеріп, ақпаратты компрессиялау қажеттігін ескеру қажет болады. Ал ақпаратты іздеу жүйесінде (АІЖ) АС көлемі, әдетте, көп болмайды, себебі мұндай жүйеде оған барынша шағын пәндік аяның дескрипторлары қосыла беріледі де бұл жағдайда ақпаратты компрессиялаудың қажеттігі болмайды [5].

Енді сөздік мақаланың компоненттік құрылымына қысқаша тоқталайық. Мәтінді автоматты өңдеу (МАӨ) жүйесінің нәтижелілігі көпшілік жағдайда сөздік мақалада (СМ) орын алатын лингвистикалық ақпараттың көлеміне және СМ құрылымының ұйымдасуына байланысты болып келеді. Ал лингвистикалық ақпараттың көлемі лингвистикалық бірліктерге (ЛБ) тіркеліп жазылатын синтаксистік және семантикалық сипаттамалардың жиынтықтарының көлеміне тікелей қатысты. Сөздік мақала құрылымындағы семантикалық, синтаксистіктік және морфологиялық бірліктерінің өзара қатынасы СМ-ның құрылымдық ұйымдасу ерекшелігіне байланысты болып келеді. Ал құрылымдық ұйымдасудың өзі СМ-да орын алатын ақпаратқа және сол мақала құрылымына тәуелді болып келетінін де ескеру қажет. Бұл мәселелердің шешімін жекелей табу мүмкін емес. Себебі ол сипаттаудың толықтығына СМ-ды құрастыру тәсілдерінің қайсыбірінің мақсатқа сай келуіне байланысты болып келеді.

Келесі қарастыратынымыз ақпарат құрылымы мен оның көлемінің ерекшеліктеріне қарай, СМ-ның типологиясына қатысты мәселелер. Сөздік мақалада алғашқы болып кездесетін лингвистикалық бірліктердің (ЛБ) параметрлерінің жиынтығын бейнелеу (көрсету) үшін әртүрлі МАӨ жүйелерінде жүзеге асқан СМ-ның үш түрлі типін ұсынуға болады: қатаң түрде ұйымдасқан СМ; шекаралары жылжымалы СМ; иерархиялық құрылымды СМ.

Кез келген сөз табына қатысты ЛБ үшін сипаттаудың барлық параметрлерін өз құрамына қосып алу алдын ала ескерілген, қатаң бірыңғайлықтағы және тұрақты ұзындыққа ие құрылым ЛБ-ның семантикалық өлшемдерімен бірге барлық сипаттауларды қамти алады [6]. Алдын ала қатаң түрде ұйымдасқан семантикалық ақпарат жекеленген АС арқылы енгізіледі [7]. Мұндай қатаң ұйымдасқан СМ-ға мысал ретінде жалпы көлемі 256 байтқа тең SKSTRAN жүйесіндегі мақаланы келтіруге болады. Бұл жүйенің АС-ы негіз сөздер сөздігі ретінде ұйымдасып, оның әрбір СМ-ы мынадай ақпаратты өзінде сақтайды:

- Y – ЛАБ-қа енгізу кезінде (жаңадан қосу, түзету, өшіру) өңдеу тәсілін көрсете отыра лингвистикалық бірлікті (ЛБ-ны) сипаттау;
- A – базалық машиналық негізді сипаттау;
- B – септік, көптік жалғаулар және т.б. жайлы морфологиялық ақпарат;
- C – меңгеруге қатысты сипаттауларды қамтитын синтаксистік шартты белгілер (кодтар);
- D – семантикалық кодтар;
- E – қосымша сөздердің аудармасына қатысты кодтар;

- F – балама сөздердің ерекшеліктерін сипаттауды қамтитын негізгі аудармалар.

Егер СМ-ның орналасқан қалпы алдын ала анықталған болса, аталған әдіс ондағы ақпаратты толықтырып және түзетулер енгізіп отыруға аса қолайлы. Бірақ кез келген жаңа ақпаратты енгізу алдын ала ескерілмеген болса, ол СМ-ды және АС-ның өзін түбегейлі өзгертуге мәжбүр етеді де құрылымдық жағынан қолайсыздыққа әкеледі. Себебі СМ-да пайдалану алгоритмінің құрастырылу-құрастырылмауына қарамай, барлық мүмкін болатын синтаксистік сипаттамалар тіркеліп, жұмыс барысында ескерілуі қажет. Мұндай сөздік мақалалардың алдын ала белгіленген орналасу қалпының көптеген жерлері толтырылмай бос тұрады. Бұл жағдай, әсіресе машиналық аударманың өндірістік жүйесінде жиі кездесіп отырады. Сонымен бірге қатаң құрылымды әрбір СМ компьютер жадынан көп орын алады, ал бұл, бір жағынан, автоматты сөздікті (АС-ты) орналастыруға қиындық тудырса, екінші жағынан, жан-жақты толық сипатталған синтаксистік ақпаратты жарыса қайталануларына да әкеп соғады.

Сөздік мақала шекарасының «жылжымалы» болып келетін жағдайында да СМ-ды қалыптастыру бірыңғай сызба бойынша жүзеге асады [5]. Мұндай сызба лингвистикалық бірлік (ЛБ) жайлы барлық ақпаратты, мәліметтің нақты типіне байланысты, он түрлі зонаға бөліп қарастырғанды жөн санайды. Әрбір зонадағы бір-бірінен ажыратылатын белгілердің саны ЛБ-да сипатталатын сөз таптарының санына қатысты болады. Сөздік мақала құрылымындағы әрбір белгінің орналасқан қалпы, оған арнайы таңылатын нөмірлермен ажыратылады. СМ құрылымында тек сәйкес нөмірмен толтырылған орын қалыптары ғана белгіленеді. Мұндай мақала барынша шағын көрініс табады, бірақ мұнда да ақпараттың екінші көшірмесі сақталып отырады. Сонымен бірге мұндай СМ-да «жылжымалы» шекаралар тек СМ-ды толтыру кезеңінде ғана орын алады да, ал автоматтың жадына сақтау тетігіне енгізу кезінде ол СМ-ның 1-ші типі сияқты қатаң құрылымға айналуы қажет. Яғни ақпаратты қосымша мәліметтермен толықтыру және оған тиісті өзгерістер енгізуге қатысты барлық мәселелер алдын ала анықталған сипаттаудың толықтығына байланысты болып келеді.

Ақпарттың бір бөлігін жеке файл ретінде жазып шығаруға сәйкес келетін СМ-ның иерархиялық құрылымында, оларды мәтінді автоматты өңдеу (МАӨ) жүйесінде қолдану тұрғысынан алғанда, синтаксистік сипаттаудың сипаттамаларына алдын ала иерархиялық орналастыру жүргізіледі. Мақаланың шекарасының «жылжымалы» болып келетін жағдайында морфологиялық және лексика-грамматикалық сипаттамалардың иерархиясы белгілі болғанымен, ЛБ-ны сипаттау деңгейі мен оған сәйкес келетін алгоритмнің арасындағы қатынас (байланыс) әлі де тұрақсыз болып келеді [8]. Егер мұндай иерархия жүзеге асатын болса, СМ біртұтастықтан қалады да, оның ішінен өзара байланыста болатын бірнеше деңгейлерді бөліп қарастыруға болады. Әрбір сипаттау деңгейі талдаудың нақты мәселелерін шешуге қатысты болады. Лексикалық сипаттау деңгейі ЛБ-ның өзін және осы ЛБ-ның СМ-ында тікелей жазылған жоғарғы деңгей сипаттамасын өз

құрамында сақтайды. Мұндай сипаттамаларға лингвистикалық бірлікке қатысты сөз таптарының белгілері, омонимия типтері, сөйлемге алғашқы синтаксистік (синтагматикалық) талдау жүргізу үшін қажетті морфологиялық сипаттамаларды жатқызуға болады .

Барынша толық синтаксистік сипаттау жеке автоматты сөздік түрінде қалыптасады да, ал оның СМ-ы ең аз мөлшердегі ақпаратты қайталауға сай, лексикалық деңгейдің СМ-ымен өзара байланыста болады.

Автоматты сөздіктегі әрбір лингвистикалық бірлік морфологиялық, синтаксистік және семантикалық тұрғыда сипатталатыны мәлім. СМ-ның типологиясын ескере отырып, ЛБ-ны сипаттау қиынға соқпауы үшін шартты түрде ЛБ сипаттамасының ең негізгі екі тобын жеке алып қарастырайық. Олар **макросинтаксистік** және **семантикалық** сипаттамалар деп аталады. Макросинтаксистік сипаттамалар ЛБ-ның лексика-граматикалық, морфологиялық және шын мәніндегі синтаксистік белгілерді, сонымен бірге мәтіндік сөзформаның мағынасын және оның сөйлемдегі қызметін анықтайды. Яғни мұндай ақпараттың болуы «кіріс сөйлеміне» синтаксистік талдау жүргізуге мүмкіндік жасайды.

Нақты лингвистикалық бірліктің макросинтаксистік сипаттауының толықтығы, оның қандай сөз топтарына жататындығымен тығыз байланысты, яғни өз тарапынан тілдің типіне қатысты сипаттаудың мүмкін болатын параметрлер (өлшемдер) санымен де тығыз байланыста болады деген сөз. Бұл байланыстылық макросинтаксистік сипаттаудың барлық құрауыш бөліктері деңгейінде де байқалып отырады.

Морфологиялық сипаттау кезінде бір тілге қатысты, бірақ әртүрлі сөз таптарына жататын және бір сөз табына жататын, бірақ түрлі тілдерге қатысты лингвистикалық бірліктерде морфологиялық сипаттамалардың саны да әртүрлі болады. Мысалы, зат есімдер мен сын есімдерде морфологиялық жекеше, көпше, септік категориялары орыс және неміс тілдері үшін ортақ болып келеді. Ағылшын тіліндегі, қазақ тіліндегі зат есімдер мен сын есімдерде «род» категориясы жоқ, ал француз тілінде септік категориясы тек жіктеу есімдік-проклитиктерге (екпін түспейтін) ғана сілтеме жасалуы мүмкін. Бұл тілдерде шақ категориясы етістіктерге ғана тән болып келеді.

Сипаттаулардың лексика-грамматикалық комбинаторикасы да тілдік бірліктің қай сөз табына және тілдердің қай типіне жататындығына байланысты болып келеді. Дәл осындай тәуелділік синтаксистік сипаттауларға да қатысты, яғни лингвистикалық бірліктің сөйлемдегі «тәртібі», оның қай сөз табы мен тілдер типіне қатыстылығына байланысты болады. Осындай функционалдық көзқараспен әртүрлі тілдерде сөз таптары мен көмекші сөздер т.б. өздеріне ғана тәндік сипатта таптастырылады.

Сонымен, лингвистикалық бірлікті макросинтаксистік сипаттау идиоэтникалық болады да, ал оның құрылымын бір ғана әмбебаптық сызба арқылы зерделеу мүмкін болмайды. Демек, нақты лингвистикалық бірлік үшін лингвистикалық ақпараттық базаны (ЛАБ) жүзеге асырарда сөздік мақаланың құрылымы өзгеріске ұшырауы керек немесе СМ-дағы ЛБ-ны сипаттау үшін екі жақты да қанағаттандыра алатын шешім табылуы керек.

Лингвистикалық бірлікті семантикалық сипаттау – лингвистикалық ақпарат базасының аса бір қажет бөлігі. Бірақ та, әзірше, оның қаншалықты толық екендігі және қалай толық сипатта құрастырылу керектігі анық емес. Қазіргі кезде семантиканың тереңдетілген жүйелері кішігірім сөздіктер үшін жүзеге асты деуге болады. Мәселен, тереңдетілген семантикалық сипаттау тек етістіктерге тіркелген Вудстың жүйесінде 3500 сөздік мақала бар, Уилкстің жүйесінде – 600, Серкон мен Виноград жүйелерінде 200-ден сөздік мақала тіркелген [9].

Егер қолданыстағы мәтінді автоматты өңдеу жүйелерін қарастыратын болсақ, мысалы, өндірістік машиналық аударма жүйелерін алатын болсақ, онда, ең жақсы деген жағдайда, семантикалық сипаттау лексикалық сипаттау деңгейінде ғана іске қосылғанына көз жеткізуге болады. Әзірше, тереңдетілген «синтаксистік» семантика жүйелері тек тәжірибе жүзінде ғана әрекет етуде.

Сонымен, өндірісте қолдануға бағытталған лингвистикалық ақпарат базасындағы семантикалық сипаттауды ұйымдастырғанда семантикалық талдау мен оған сәйкес семантикалық сипаттауды, тек синтаксистік талдау құралдарының тиімділігі өте төмен болған жағдайда ғана іске қосуға болады. Алдын ала зерттелген алгоритмдердің мүмкіндіктерін ескеретін болсақ, семантикалық сипаттау да рет-ретімен қалыптасады. Осындай тәсіл бойынша семантикалық сипаттау жүйесі ашық жүйе түрінде қалады және лингвистикалық ақпарат базасының (ЛАБ-тың) сөздік жүйесін қайта құрастырмай-ақ оны тек толықтырып тұруға әбден болады. Сонымен бірге сөздікпен қамтамасыз ету құрылымын орнатуда, яғни ЛАБ-тың ішкі сызбасын таңдауда, компьютердің жадына жазу деңгейінде сөздік мақаланың (СМ-ның) ерекшеліктерін анықтап алу аса маңызды [10].

ӘДЕБИЕТ

1. *Гончаренко В.В., Добрускина Э.М., Котельникова Н.М. и др.* Автоматическая переработка текста. – Кишнев, 1978.

2. *Штерн И.Б.* Об одном способе построения русского морфемного словаря для машинного перевода. – НТИ, 1966, №11.

3. *Кулагина О.С.* Исследования по машинному переводу. – М., 1979.

4. *Марчук Ю.Н.* Вычислительная лексикография. – М., 1976.

5. *Аполлонская Т.А.* Организация словарной статьи в многоуровневой системе машинного перевода. – В кн.: Инженерная лингвистика и преподавание иностранных языков с помощью ТСО. – Л., 1981.

6. *Шаляпина З.М.* Англо-русский многаспектный автоматический словарь (АРМАС) // Машинный перевод и прикладная лингвистика. Вып. 17. – М., 1974.

7. *Леонтьева Н.Н., Кудряшова И.М., Соколова Е.Г.* Семантическая словарная статья в системе ФРАП. – М., 1979.

8. *Аполлонская Т.А., Соколова С.В.* Уровневое построение алгоритма машинного перевода. – В кн.: Инженерная лингвистика и преподавание иностранных языков с помощью ТСО. – Л., 1981.

9. *Чижаковский В.А., Беляева Л.Н.* Тезаурус в системах автоматической переработки текста. – Кишинев, 1983.

10. *Cercone N., Goebel R.* Approaches to Knowledge Representation. – In: Empirical Semantics. – Bochum., 1981.

ҚАЗАҚ СӨЗІНІҢ «ТІЛ – ҚАЗЫНАСЫ»

«Болашақ». Республикалық қоғамдық-саяси
жастар газеті. №18(81). Сәрсенбі, 8 маусым 2011 жыл.

1969 жылы Қазақстан Ғылым академиясы Тіл білімі институтында «Статистическое и информационное изучение тюркских языков» деген атпен Бүкілодақтық ғылыми жиынның өтуі қазақ тіл білімі үшін елеулі құбылыс деуге болады. Осы кезеңнен қазақ тілін статистикалық зерттеу мен автоматтандыру әдістері Тіл білімі институтында жүйелі сипатқа ие бола бастады.

«М.О.Әуезовтің «Абай жолы» романының жиілік сөздіктері» және «М.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері» атты еңбектер қазақ тілі мәтіндерін статистика тәсілімен зерттеудің және лексикография саласын автоматтандырудың алғашқы нәтижелері деп қарастырған жөн. Жазушы тілінің лексикалық байлығын, автордың сөз қолданудағы стильдік тәсілінің өзіндік ерекшелігін, қазақ тілінің көркемдік және бейнелеуіш мүмкіндіктерін меңгеру деңгейін осы аталған лексикографиялық еңбектер бойынша бағалауға зор мүмкіндік туды.

Электронды есептеу машиналарының (ЭЕМ) көмегімен ұлттық тілдерді зерттеу мен одан туындайтын ақпараттарды автоматты түрде танып-білудің негізінде өткен ғасырдың 80-жылдарынан бастап тілдердің машиналық қорын жасау жұмысы туралы ой-пікірлер жүзеге аса бастады деуге болады. Аталған мәселеге қатысты отандық ғылыми жиындарда орыс, украин, грузин, эстон тілдеріне қатысты баяндамалардың қатарында қазақ тілінің машиналық қорын жасау мәселесі де арнайы тыңдалды. Мысалы, 1988 жылдың мамыр айында Мәскеуде Кеңес Одағы түркітанушылар комитетінің XIV Пленумы ашылып, оның күн тәртібіне түркі тілдерінің, оның ішінде қазақ тілінің машиналық қорын жасау мәселесі арнайы қойылды.

Сонымен бірге, қазақ сөзінің «Тіл – қазына» атты деректер базасы үлкен көлемді мәтіндер мен көп көлемді сөздіктерге қатысты ақпараттық-сұраныс жүйесі түрінде көрініс табуы да қазақ мәтіндер корпусының компьютерлік базасын құрудың мақсат-міндетіне айналып отыр.

Көптеген ғалымдардың пікірлерін қорыта келе, біз қазақ сөзінің «Тіл – қазына» атты деректер базасы жүйесінің түсінігі – көп тармақты әрі терең

автоматтандырылған жүйені аңғартады демекпіз және оның атқаратын міндеті ақпараттық барлау мен тіл білімі саласындағы (жалпы филология бағыттағы) зерттеушілік болуы тиіс.

Қазақ тілі жөніндегі деректерді жинаудың, сақтаудың, талдаудың және салыстыра зерттеудің жаңа әдістеріне көшу, сондай-ақ лингвистикалық дерек көздерінің автоматтандырылған сөздіктер мен грамматикалар тәрізді жаңа түрлерін игеру ісі өмірге жанасымды және тиімді екені сөзсіз. Әрине, бұл үшін біз жалпыфилологиялық дәстүрлер мен мәдениетке, тілді танып-білуге, практикалық жаңа міндеттер тәжірибелерін мұқият ескеруге тиіспіз.

Қазақ сөзінің «Тіл – қазына» атты деректер базасының лексикалық сипаты түпнұсқа мәтіндерінің типіне байланысты деп санай отыра, қазақ тілінің жүйесін танытатын сөйленіс жағдаятына: көркем әдебиет тілін, газет-журнал тілін, ғылым мен техника тілін, ұйымдастыру-нұсқау құжаттары тілін, заң орындары мен сот мекемелер тілін, оқу орындары тілін, сауда, дәріхана, емханалардағы және т.б. сөйленіс түрлерін жатқызуды қолдар едік.

Келесі бір қомақты мәтіндік бірлік – ол терминдер мәселесі. Қазіргі ғылыми-техникалық және іскағаздары мәтінінің негізгі өзегі терминдер екені айқын. Жалпы сөйлеу тілінің аясындағы сөздер емес, терминдердің тілдік қоршауы (мәнмәтіні) өзгеше бір тілдік жүйе, басқаша айтқанда, терминдер өрісі болып табылады. Осы аяда өмір сүретін термин үшін барлық ғылыми шығармалардың сөзқолданыс машықтары етене, ортақ болып келеді.

Халық тілі сол халықтың жанды қазынасы болғандықтан, соның бүкіл болмысын барлық ерекшеліктерімен қоса қамту қазақ сөзінің «Тіл – қазына» атты деректер базасының міндеті деп түсіну қажет сияқты.

Жоғарыда сөз болған,1988 жылы Мәскеуде өткен Пленум мәжілісінде түркі тілдерінің көпаспектілі машиналық қорын жасау қолданбалы ілімдердің мұқтаждығына сай жүзеге асырылуы тиіс екендігі баса айтылған болатын. Өкінішке орай, мұндай жақсы бастама сөз жүзінде ғана қалып, түркі тілдерінің көпаспектілі машиналық қорын өмірге келтіру әлі де болашақтың ісіне айналып отыр.

Ал қазіргі жағдайда, ең алдымен, егеменді еліміздің мемлекеттік тіліне айналған ана тіліміз – қазақ сөзінің «Тіл – қазына» атты деректер базасын ғылыми тұрғыда құру – жаңаша ойлап, жаңаша шешудің ең тиімді жолы деуге әбден-ақ болады.

Біз, осы айтылғандарға қосыла келе, қорытынды ретінде айтқымыз келгені мынау еді: қазақ сөзінің «Тіл – қазына» атты деректер базасы зерттеуші адамның қалып-қабілетін әлденеше есе арттырумен бірге ізденушінің шығармашылық қуат көздерін аша түсетіні айғақ, сөйтіп бұл жаңа мүмкіндіктер қазақ тілінің жүйелілік қасиеттерін жетілдіруге және тіл жүйесін мұқият тануға жұмсалады деп тұжырымдауға әбден болады. Әрине, бұл аталып отырған қазақ сөзінің «Тіл – қазына» атты деректер базасын толық сипаттап, өмірге келтіру үшін оның жоба-жолдары тек А.Байтұрсынұлы атындағы Тіл білімі институты басшылығы мен ғылыми кеңесінің деңгейінде ғана емес, ол мемлекеттік дәрежеде, яғни ҚР Білім және ғылым министрлігі, Ғылым комитеті, Тіл комитеті деңгейінде қаралуы қажет.

Себебі, мұндай жалпы филологиялық, ұлттық және мәдениет дәрежесіндегі күрделі жүйені іске қосу – қажетті мамандарды, мысалы, жоғары дәрежелі инженер-бағдарламашыларды, компьютерлік лингвистика саласының мамандарын және тиісті қаражат мүмкіндігін шешуді талап етеді.

THE FOUNDER OF THE KAZAKH LINGUISTICS AND THE VICTIM STALIN'S REPRESSIONS

(tragic fate of Professor Khudaibergen Kuanovich Zhubanov)

CENTRAL EURASIAN STUDIES: PAST, PRESENT AND FUTURE. 17-19 mart 2009. T.C.Maltepe universitesi, University of Tokyo, University of Tsukuba ISBN: 978-975-6760-44-4. T.C.Maltepe universitesi YAYINLARI; 46 (2011/5). P. 429-435.

Kazakhstan, Almaty, Kh.K.Zhubanov *– doctor of philological science, professor, leading researcher of the Institute of linguistics named after A.Baitursynov, Ministry of education and science, the science committee of the Republic of Kazakhstan.*

The theme of terrible repressions by the end of 30-th years is an inexhaustible. Also that time was terrible for Turkic people, their languages, literature, history and culture who were a part of the USSR.

In 1937 there was still an opposition of the academicians – communists and the scientists, who were not party-members, the latter ones were more numerous. There were two categories of party-members at the Academy – the scientists who joined the party in the soviet period of time, mainly in the 30th and state and party functionaries, usually having long years of party membership, "appointed" to work in a scientific field and taking it as a lower position.

In August and October of 1937 a big group of officials and researchers suffered from repressions. They worked in the Institute of the Oriental Studies at the Academy of Sciences of the USSR. Many of them studied the Far East, Japan, China, Korea, Mongolia, Manchuria and Tibet.

Losses among the national intelligentsia in Kazakhstan were more considerable than in other countries. There were two waves of repressions: in 1929-1931 and in 1937-1938 years.

During the years of the Civil war the position of national intelligentsia was quite different and many of it's representatives stood apart of the politics. In Kazakhstan intelligentsia, being less numerous, joined together to the national party "Alash". This party workers and the workers of the national government "Alash Orda", set in 1917-1919, fought with communists, but in 1919-1920 the majority of them accepted the soviets and even for some time were members of Kazakhstan government. But the state leaders didn't believe "Alash Orda" and soon they were dismissed from their positions. In 1929-1931 a campaign was

launched against those ones, who earlier were “Alash Orda” members, many of them were arrested, some of them were shot and the others spent several years in the camps and exile.

In 1937-1938 there was even more severe wave of repressions. All the prominent “Alash” members were killed, all the distinguished Kazakh communists perished. The elder generation of national culture workers was eliminated almost altogether. Among them were people, developed into intellectuals before the revolution as well as those ones, belonging to the first soviet generation of Kazakh intelligentsia [1]. The youngest generation of Kazakh intelligentsia being grown up during the years of the soviet power, suffered less. They began their activity in the 30th.

In the 50th the process of rehabilitation in Kazakhstan was incomplete and concerned only the killed communists, although in many soviet republics the most of scientists and cultural workers were justified.

Prominent cultural workers, connected with “Alash” and “Alash Orda”, were rehabilitated only in 1988-1992. Now in Kazakhstan their names are especially honoured.

“Alash” members didn’t recognize the soviet government and in 1917 organized “Alash Orda”, i.e. their own government headed by A. Bukeikhanov, A. Baitursinov was one of it’s members.

Baitursinov Akhmet was born in 1870 in Kustanai region. He had secondary education and wasn’t a party member, he worked as a teacher at the state Kazakh University and was a scientific and a literary worker. On the 12th of June 1929 Baitursinov was arrested and put to trial (criminal case 58-7,58-11 and 59-3 RSFSR Criminal Code). He was accused as one of the founders and leaders of a counter-revolutionary nationalist organization “Alash” and of the “Alash Orda” government. As the document about the sentence execution tells (p.105) Akhmet Baitursinov was executed on the 8th of December 1937 [1].

The name of Akhmet Baitursinov in Kazakhstan became not only often mentioned, but in many ways officially canonized, especially since the declaration of independence had been announced in 1991. One of central Almaty streets and the Institute of Linguistics of the Kazakh National Academy were named after him.

In this report we would like to tell in short about a man of great talents, about the most appreciable person after A.Baitursynov, the famous Kazakh linguist Khudaibergen Zhubanov.

As we know, in 1937 the first professor, the well-known scientist-linguist Khudaibergen Kuanovich Zhubanov also became the victim of lawlessness. When we say about the origin and formation of the Kazakh soviet linguistics we should remember with a deep respect and gratitude people, who devoted all their lives to native linguistics. One of these people was professor Khudaibergen Kuanovich Zhubanov who successfully worked for the Kazakh linguistics in the 20-th and 30-th years of the last century. He is considered one of the founders of the Kazakh soviet linguistics. Khudaibergen Kuanovich Zhubanov was the first researcher in the history of Kazakh linguistics who was awarded the title of a professor [2; 3].

The last publication of the book named “The history of the Kazakh SSR. Era of socialism” gave a thorough description of the condition of science and education in the Republic in 30-th years. We’re reading a short speech: “In the years of the second five-year period the first pedagogical Institute – Kazakh pedagogical University named after Abay – changed into a large educational Institute and an important cultural hearth… The scientists studied and developed weighty problems of literary criticism, linguistics and the history of the Kazakh SSR there. Kazakh orthography and a new corresponding to the Kazakh language the written language was developed by Zhubanov. Professor Khudaibergen Kuanovich Zhubanov wrote the first textbooks, researches in grammar, dialectology and history of the Kazakh language” [4].

Although the language of historical chronicle is very stingy, but these facts can show Zhubanov’s person, as a prominent linguist. Serving to his native people became the meaning of his life.

K.K.Zhubanov was born in 1899 in Kusuaktum place of Temir-Urkachevsky volost, Temir district of Uralsky region (now Aktobe region) in the family of a middle peasant Kuan Zhubanov.

In 1914 Zhubanov studied in Medrese “Khusainiya” in Orenburg. He learned there secular science and the Russian language. There Khudaibergen found friends, one of them was Shamgaly Kharesovich Saribaev – one of the founders of the theoretical course of the native language methods of teaching. In 1916, he finished the fourth class of the Dzhurunsk primary school. In 1918 Khudaibergen Kuanovich Zhubanov graduated with honours from Ilek’s college, two classes of the college (the 5th and the 6th) he finished in a year. Thus he began from Koran in “Ospan-Ishana” and finished the Ilek’s Russian secondary school. It should be mentioned that K.K.Zhubanov had got knowledge during his early stage of studies that helped him later on to become a good linguist. During those years he had been learning the languages with different structures: Arabic, Persian and Russian. At the age of 18 Zhubanov could write and speak Turkic, Arabic and Persian.

Zhubanov was interested in literature from his childhood. Knowledge of the oriental languages opened a new world before him. He read in original a lot of Arabic and Persian literature, including “Shahname” of Firdausi, “Gulistan” of Saadi, “Leila and Medzhnum” of Hisami, etc.

Later, when Zhubanov became the professor of the Kazakh language Department in Kazakh pedagogical Institute named after Abay he used his knowledge; he took an active part in all events dedicated to different problems of culture and art of Central Asia and Kazakhstan.

In 1925 when Khudaibergen worked as an inspector-methodologist he was actively learning three languages at the same time: German, English and French, and he was interested in psychology, logic and reflexology. He studied the works and articles of academic I.P.Pavlov, V.M.Bekhterev and other famous Russian scientists. At the same time he corresponded with many linguists from Moscow, Kazan, Baku, Tashkent, Leningrad, such as E.D.Polivanov, B.A.Bogoroditski, I.I.Meshaninov, Agamaly-Ogly and others.

In the spring of 1928, Zhubanov was called to Kyzyl-Orda. At that time Kyzyl-Orda was the capital of Kazakhstan. Zhubanov was an inspector of Narcompros. In January of 1929 he was sent to Leningrad Institute of the Oriental languages to the post-graduate courses in Turkology, and then a year later he continued his studies at the post-gratuate courses of the Yafety Institute of the USSR Academy of Sciences.

At the same time K.K.Zhubanov on his own was engaged in research work in the field of general phylology and Turkic languages phylology.

Besides Khudaibergen Kuanovich Zhubanov took an active part in the developing of a new Latin – based alphabet.

In June of 1929 Zhubanov made a speech at the scientific-orthographical conference. This speech was devoted to the development of the main principle of Kazakh orthography.

In 1930, Zhubanov made a speech on the first history and economy congress of Kazakhstan in Almaty. The theme of his report was “The meaning of meaningless words”, that lecture became the result of solid scientific research of several Turkic languages. The report “Cattle-breeding interjections of the Kazakh” was devoted to the analyses of Kazakh interjections addressed to domestic animals.

According to some data in 1937 some more scientific works by K.K.Zhubanov had to be published, among them: “Materials to the scientific grammar” (volume 1, phonetics) «One more time about the verb “de”» and others. At present their destiny is unknown.

Zhubanov was not only an expert of Turkic and other languages, but also he was an outstanding publicist. His wonderful work “Kui” distinguishly and originally formulated the question and used the wealth of actual materials [5]. As a whole in this work the author made more precise the sequence of appearance of national musical genres by the way of systematization and studying the historic-linguistic facts from Turkic culture and languages and by the attraction of famous works such as Abdul-Khadir Muragi, Rauth Ekta, Al-Farabi, Kizevetter, Belyaev, Uspenskyi, professor Phitrat , professor Oganezashvilli and others.

Zhubanov in his publicistic works wrote many interesting and cognitive facts from the classical written literature. His article about the language of Abay isn’t of less significance. It is devoted to Abay – the classic of Kazakh written literature and is very interesting and substantial. This article is a valuable contribution to the Kazakh history of literature. He was the first researcher who described Abay’s language, his innovation, strong skills and the original character of his versification. As an excellent expert of a treasure house of the world culture, Zhubanov researched the past. He would like to show the place and meaning of Abay Kunanbayev in the Kazakh history of culture and literature.

Khudaibergen Kuanovich Zhubanov is rightly considered a prominent researcher of the Turkic languages.

Being the post-graduate student of the Leningrade East Institute and then the Yapety Institute of the USSR Academy of Sciences Zhubanov showed his outstanding gifts in the oriental studies and of course as a shrewd observer of national culture. When he was a head of the Kazakh language department in the

Kazakh pedagogical Institute named after Abay, he tried to learn the peculiarities of runic inscriptions. He handled the language of ancient Turkic monuments and his knowledge was helpful. K.K.Zhubanov skilfully combined his pedagogical activities and scientific research work devoted to the problems of Kazakh and other Turkic languages. He wrote the whole series of valuable scientific works and they are still urgent at present. The comparative-historical method in linguistics characterizes his works.

Really scientific analysis of linguistic phenomena is impossible without a historical approach. A kind of enthusiasm for fashionable at that time linguistic trend-yafetology can't depreciate K.K.Zhubanov's theoretical heritage.

K.Zhubanov's work, devoted to auxiliary and complex verbs [6], showed him as an expert of lexical and grammatical nuances of his native language. This Zhubanov's work is interesting to the language historians not only because of the research of part of the sentence combination but also it attracts attention to the etimological research of grammatical and lexical forms of several Turkic languages. This work shows the expediency of the comparative-historical method for Turkic languages reaserch.

Especially we would like to mention his textbook about Kazakh grammar. It is the first textbook for the pupils of the 5-th and 6-th classes of a secondary school [7]. That book became the main model for other grammatical textbooks. It was one of the first textbooks for the Kazakh secondary school. This grammar textbook gave a system of main problems of Kazakh language word structure and phonetics.

The scientist gave his view on the problem of the qualitative and quantitave structure of Kazakh phonemes. Professor K.K.Zhubanov came to the conclusion that many phonetic phenomena such as junction and displacement of syllables, the stress in a sentence and especially syngarmonical laws contribute strongly to the firm connection of word elements between themselves. He based exactly on the Kazakh phonology statements all his further researches devoted to the problems of alphabet, orphography and other branches of applied linguistics.

Besides grammar K.K.Zhubanov in the same 1936 published the Kazakh language grammar for the 5th - 7th forms of the secondary school. This programme is one of the first documents, showing the transformation of the Kazakh language to the subject of scientific research. It showed to the full extent K.K.Zhubanov's view of a complex sentence syntax.

All this shows that K.K.Zhubanov takes by right a worthy place in the history of the Kazakh school as a prominent teacher.

Zhubanov's textbook took a high appreciation among teachers' society of our Republic in his time. Famous soviet turkolog S.E.Malov appreciated it greatly. In 1935 there was a congress dedicated to cultural building, and there Malov said that the project done by Zhubanov is extremely scientific and principally consistent [8]. But in the period of the personality cult the Kazakh language grammar by K.K.Zhubanov was wrongly criticized, there appeared articles searching for sabotage traces in grammar statements. The whole series of K.K.Zhubanov's articles are devoted to the problems of practical linguistics – orphography, terminology, methods of the Kazakh language teaching and etc. Original character

of language phenomena researches and his activities in the field of education of future linguists give every reason to state that K.K.Zhubanov was one of the founders of Kazakh soviet linguistics.

The name of professor Khudaibergen Kuanovich Zhubanov was popular at that time with the Kazakh intelligentsia especially with education workers. Moreover, he was well known among distinguished Russian Soviet scientists such as academicians I.I.Meshaninov, A.N.Samoilovich, V.V.Bartold, S.E.Malov, K.K. Udakhin, A.C. Borovkov, etc.

Alongside with scientific – educational activities K.K.Zhubanov took an active part in public activities. He took the initiative to organise scientific conferences, devoted to the outstanding figures of the past – Firdoussi, Abay and he made an interesting report there. In 1935 he was elected a chairman of organizing committee at the cultural workers congress and then he was elected a delegate from Kazakhstan at the XVI all-Russia Congress of Soviets.

It is necessary to recognize that in 1937 there were only few professional Kazakh-linguists and almost all of them were done away with (with the exception of the young researchers) by Stalin's repressions. Moreover, because of those events Kazakh linguistics lost its founders. Stalin's punitive machine worked down precisely and without any extra fuss.

When Zhubanov was arrested and accused of nationalism he was only 38 years old, he was full of creative power. The original textbook of grammar, was published in 1936, became the main cause of accusation. Professor used materials from the works of the first founders of "Alash Orda": Seifullin, Mailin, Zhansugurov, who were declared enemies at the beginning of 1937.

Zhubanov supported Akhmet Baitursynov's principles of the Kazakh orthography, who was claimed as a counter – revolutionary. All these and other curcumstances were the foundation to declare Zhubanov an enemy. I'd like to tell about some other facts referring to the last days of the scientist in prison. One of such facts, showing firmness and purposefulness of Zhubanov's spirit, refer to the fragments of his speech, recorded in "Criminal Case №11020", which accused him. This case kept in the archives of the State Security Commitee of Kazakhstan. «The archivist-investigatory case №9555387» is kept in the same place, this case says that «it is opened on the 19th of November 1937» and «is completed on the 23^{d} of December 1937» [9, 102-107]. In the examination record were some Zhubanov's phrases having deep meaning: "Our leaders cried about enemy machinations of rich people in our country. They cried about rich people's (the khans) resistance to the socialist rearrangement of Kazakhstan, tried to fight the rich with the help of the "activists". They forgot that the rich (the khans) were the glory and the **strenghth** of the Kazakh people, they forgot that the khans gave to the state cattle, meat, wool, leather and bread. And when the institution of the khans was ruined, when all the wealth of our country was **ruined utterly** and the centuries-built economy of Kazakhstan was destroyed, then being no better off than at the start, we began to cry out about exaggerations and their consequences and the necessity of improving the faults".

"There were many excesses in the presence of Goloshekin. It meant a death

of a considerable part of Kazakh people. All population of our country lost their cattle. It was confiscated by the Soviet power. The whole nation was put to elimination. Because of those events there were huge amount of innocent victims. However, Russian people in Kazakhstan did not starve to death…".

Further the scientist in his discussions looked for the reason of such a tragic fate of all the Kazakh people. We can see that from his further speech: "...Improving now the extremes, we judge again the same Kazakh workers. In that and in this way it causes the loss of the Kazakh workers. In such a situation dying out of the Kazakh people is inevitable. We judge people for squandering money, and why we don't judge people for squandering a whole nation, isn't it a crime? The Soviet government knows about that events but does not pay attention. Therefore, Goloshekin is not an only figure in this process. The main purpose of the Soviet power is building of socialism and that building would proceed if they lose the whole nations. Thus, Kazakh people have become the victim of the faults done while rebuilding the economy. Our purpose is to protect the nation from extinction". The last speech of Zhubanov before the execution became prophetic for Kazakh people: "Our culture cannot be national, in the presence of the soviet government it will be russified. If we want that our people won't become the victim of socialism we should free them from socialism. However, it will be done by the way of establishing an independent bourgeois-democratic government".

Thus we can say that ideas and suggestions made by the founder of the Kazakh linguistics, the first Kazakh professor Kh.K.Zhubanov harmonise greatly with the present situation of the Kazakh people.

Zhubanov personaly could not take part in "Alash Orda" because of his age. His case says (p.201): "Even since 1917 Zhubanov took an active part in publishing the Kazakh nationalist magazine "Tez" and then being a member of nationalist group, published a nationalist magazine "Kizgaldak" [1]. As we know, he was accused according to the articles number 58-2, 58-8, 58-9 and 58-11. The first article accused Zhubanov of espionage, because he was interested in a Japanese language. Then, the final verdict changed to the other three articles before his death penalty. Khudaibergen Kuanovich Zhubanov was executed on the 25-th of February in 1938.

Zhubanov was married and had six children. The eldest child was 12 years old and the youngest one was three months old. His wife Raushan brought up six children by herself. Afterwards these children graduated from the university and continued their father's work in philological science.

Zhubanov was not accused for belonging to "Alash Orda", that fact did not sae him from death, but it helped him to be rehabilitated.

To bring in a verdict of not guilty for K.K.Zhubanov there were received references from candidates of philological science Makhmudov and Saikiev, doctor of philological science Balakaiev and from the manager of history of art department of the Kazakh SSR Academy of Sciences Erzhakovich.

All these scientists have a positive opinion about Zhubanov's published works and declare them to have definite scientific value, Zhubanov's works don't have any anti-party or anti-soviet ideas and statements (p.211) [1].

And then there was done the following conclusion: “The present case… place for consideration of the Military Board of the USSR Supreme Court in order to repeal a sentence to Zhubanov Khudaibergen for lack of corpus delicti” (p.212) [1]. The conclusion from 03.04.1957 was confirmed by the deputy General Procurator of the USSR and on 03.10.1957 the Military Board of the USSR Supreme Court repealed a sentence done in 1938 (p.213) [1].

Thanks to those events, government republished Zhubanov’s works. They returned to the scientific rotation earlier than A.Baitursynov’s works and his main scientific works were republished in 1966.

Kh.Zhubanov’s name, the name of the first founder of Kazakh linguistics, the researcher-turkolog was expunged from the memory of Kazakh people till October 1957. Premature loss of this phenomenal scientist with keen mind and encyclopaedic knowledge in the field of humane sciences influenced greatly further development of the Kazakh linguistics, which wasn’t solid enough in the 30-th. However, his works done within 20 years, the grateful generation could fully appreciate. The name of the professor is living in his priceless works, in the works of a new generation of scientists.

At present Zhubanov’s tradition in science is going on. His scientific works are republished. Anniversary dates are celebrated in big scientific institutions and at universities of the Republic and scientific-theoretical conferences are held, devoted to the scientist, who was the first to be honoured with a professorial rank. In the professor’s native land – in Aktobe city the State University was named after Kudaibergen Zhubanov. There was built a monument and the avenue in front of the university was named after brothers Zhubanovs – Kudaibergen Zhubanov and his own brother academician Akhmet. It is “Brothers Zhubanovs’ street”.

The fundamental linguistic researches of the main problems of the Kazakh linguistics and his scientific activity characterize Zhubanov as an eminent representative of science of his era.

References

1. *Ашнин Ф.Д., Алпатов В.М., Насилов Д.М.* Репрессированная тюркология. – М. : Вост. лит., 2002. – 296 с.

2. *Ауэзов М.О., Кенесбаев С.К.* Основоположник казахской лингвистики// Худайберген Жубанов и казахское советское языкознание. Алма-Ата: Наука, 1990, стр. 21-25.

3. *С.К.Кенесбаев.* Профессор Худайберген Куанович Жубанов // *Х.К.Жубанов.* Исследования по казахскому языку. - Алма-Ата: «Наука», 1936. – 361 с.

4. «История Казахской ССР. Эпоха социализма». Алма-Ата, 1963, стр. 529-531.

5. *Х.К. Жубанов.* К вопросу о возникновении казахского музыкального жанра – кюй. Историко-лингвистический очерк. Кзыл-Орда, 1936 (на каз. яз.).

6. *Х.К.Жубанов.* Исследования по казахскому языку. Вып. 2. Заметки о вспомогательных и сложных глаголах. Алма-Ата, 1936.

7. *Проф. Қ.Жұбанов.* Қазақ тілінің грамматикасы, I бөлім. Сөздің жалпы құрылысы (жалпы морфология). Орта мектептің V класына арналған. 1936, Қазақстан баспасы, 74 б.

8. Стенограмма выступления С.Е.Малова на съезде по культурному строительству в Казахстане. (См. ЦГА КазССР, ф. 81, оп. 3, д.1086, лл. 10-12).

9. *Жұбанов Е.Қ., Оралбай Ә.* Профессор Құдайберген Жұбановтың саяси памфлеттері // Жұбановтану жылнамасынан. – Ақтөбе, 2004. 186 б.

ÖZGEÇMİŞ

Bu bildiri 20-30 yıllarda kazak dili filolojisi alanında başarıyla çalışmalar yapmiş ve kazak dilciliği temellerini atanlardan biri olan dil bilgini, alim Hudaybergen Kıvançoğlu Jubanov'un hayatına ışık tutmak maksadıyla yapılmıştir. O, kazak filolojisi tarihinde ilk olarak profesör ünvanını almış ve 1937 senesi Stalin'in korkunç usulsüz soykırımına mağdur kalmıştır. Leningrad Doğu Enstitüsü araştırma görevlisi ve Dil bilimi enstitüsü dokora öğrencisi iken o, kendini doğu dil bilimciliği sahasında geniş bilgililiği ve titiz çalışmalarıyla kendini gösterebildi. H.K. Jubanov, özellikle Kazakistan'da Abay Pedegojik Üniversite'sinde Bölüm başkani olarak çalıştığı yıllarda verimliliğini tam anlamıyla ortaya koyabildi.

H.K. Jubanov'un imla, terminoloji, kazak dili öğretim yöntemleri ve benzeri alanlarda yaptıkları eserleri kazak dilbilimciliğinin esasları olarak seri halinde yayımlanmiştır. Dil sahasında yaptığı çalışmalarının orijinal olması, dilcilik alanında yetiştirdiği öğrencileri ve emekleri H.K Jubanov'un kazak dil bilimi temel taşı atanlardan biri olduğunu gösterir. Onu bir dil bilimci alim olarak bütün rus alimleri ile soviyet alimleri: İ.İ. Meşaninov, A. N. Samoyloviç, V.V. Bartold, S.E. Malov, K.K. Yudahin, A.K. Borovkov vb. saydı ve hürmet ettiler.

Kazak dilbilimciliği temelin oluşturuculardan, büyük dilci H.K. Jubanov'un ismi 1957 senesi ekim ayına kadar Kazak halkının hafzasından silmeye çalıştılar. Ender zeka sahibi ve yaşayan ansiklopedi olarak tanılan bu bilginin zamansız kaybı, 30. yılları daha yeni gelişmekte olan kazak dil bilimi sahasının, sosyal bilimler alaninin gelişmesinde büyük kayp olarak tanımlanabilir. 20 yıl gibi kısa süre içerisinde yaptığı verimli çalışmalarına bile yeni nesil saygı gösetermekte ve büyük hürmetle anmaktadır. Alimin isim değer biçilmez eserleriyle ve dil bilimciliği alanında yeni çalışmalar yapmaka olan neslin aklında daima kalacaktır.

S U M M A R Y

The report briefly describes the tragical destiny of a scientist Khudaibergen Kuanovich Zhubanov who worked in the 1920th and 1930th years in the field of Kazakh phylology and is highly appreciated as one of founders of Kazakh Soviet Linguistics. He was the first who was awarded the rank of a professor in the history of Kazakh philology. In 1937 he became a victim of terrible lawlessness and repression. When he was a postgraduate of Leningrad Eastern Institute (nowadays Institute of Linguistics) he discovered himself as a fine observer and investigator – a scientist of a wide range in the field of oriental studies. K.K.Zhubanov's talent and gifts were demonstrated especially brightly in the period of his teaching in the Kazakh pedagogical Institute named after Abay. He worked there for several years as a head of Kazakh language faculty.

The whole series of K.K.Zhubanov's articles are devoted to the problems of practical linguistics – orthography, terminology, methods of Kazakh language teaching etc. Original character of language phenomena studies, active work for the preparation of linguists give foundation to affirm that K.K.Zhubanov was one of the founders of Kazakh Soviet linguistics. He was known and appreciated as a remarkable linguist by such prominent Russian Soviet scientists, as academician I.I.Meschaninov, A.N.Samoilovich, V.V.Bartold, C.E.Malov, A.K.Udakhin, A.K.Borovkov and others.

The name of Khudaibergen Zhubanov – a founder of Kazakh linguistics, a prominent researcher of Turkic language was crossed out of the memory of Kazakh nation. The premature loss of this phenomenal scientist with searching mind and really encyclopedic knowledge in the field of humanitarian science influenced remarkably further process of Kazakh linguistics development. Grateful generation highly estimated his works done only within 20 years.

The professor's name continues now to live in his precious works, in researches and creations of linguists of young generation.

Р Е З Ю М Е

В докладе кратко освещается трагическая судьба ученого-лингвиста Худайбергена Куановича Жубанова, который успешно трудился в 20-ые и 30-ые годы в области казахской филологии и по праву считался одним из основоположников казахской советской лингвистики. Он первый в истории казахской филологической науки был удостоен звания профессора и в 1937 году стал жертвой страшнего беззакония и грубого произвола сталинской репрессии. Будучи аспирантом Ленинградского Восточного института, а затем Института яфетидологии (ныне Институт языкознания), он проявил себя как тонкий наблюдатель и исследователь-лингвист широкого диапазона в области востоковедения. Талант и способности Х.К.Жубанова особенно ярко проявились в период его преподавания в Казахском педагогическом институте им. Абая, где он несколько лет работал заведующим кафедрой казахского языка.

Целая серия статей Х.К.Жубанова посвящена вопросам практического языкознания – орфографии, терминологии, методике преподавания казахского языка и др. Оригинальный характер исследований языковых явлений, деятельное участие в подготовке филологических кадров дают основание утверждать, что Х.К.Жубанов был одним из основоположников казахского советского языкознания. Как замечательного ученого-лингвиста его знали лично и ценили такие видные русские советские ученые, как академики И.И.Мещанинов, А.Н.Самойлович, В.В.Бартольд, С.Е.Малов, К.К.Юдахин, А.К.Боровков и другие.

Имя Худайбергена Жубанова – основоположника казахской лингвистики, крупного исследователя-тюрколога, до октября 1957 года было вычеркнуто из памяти казахского народа. Безвременная утрата этого феноменального ученого с пытливым умом и поистине энциклопедическими знаниями в области гуманитарных дисциплин заметно сказалось на дальнейшем процессе развития не до конца еще окрепшей в 30-х годах науки о казахском языке. Но даже то, что он успел сделать за каких-то 20 лет, благодарное поколение сумело по достоинству оценить. Имя профессора ныне продолжает жить в его бесценных трудах, в поисках и творениях молодого поколения ученых-лингвистов.

МЕМЛЕКЕТТІК ТІЛДЕГІ ҰЛТТЫҚ «ТІЛ – ҚАЗЫНА» ДЕРЕКТЕР БАЗАСЫНЫҢ ТЕОРИЯЛЫҚ НЕГІЗДЕРІ

ҚР БҒМ ҒК А.Байтұрсынұлы атындағы Тіл білімі институты. «Мемлекеттік тіл – ұлттық бірегейліктің негізі». Қазақстан Республикасы тәуелсіздігінің 20 жылдығына арналған ғылыми жинақ материалдары. – Алматы: «Қазақ энциклопедиясы», 2011. 291-318-б.б.

Қазақстан Республикасы жеке егемен мемлекет ретінде Тәуелсіздігін жариялаған кезден бастап қазақ тіліне мемлекеттік тіл мәртебесі беріліп, оны қоғам мен ғылымның барлық салаларында кеңінен қолдану мәселесі тұрды. Бұл мәселені шешудің бір жолы қазақ тілін ғылым тіліне айналдыру мақсатында іс-шаралар атқарылып келеді. Соңғы жылдары ғылыми-техникалық прогрестің дамуын жеделдетуге компьютерлік технологиямен бірге барлық ақпараттық индустрия да өз әсерін тигізе бастады. Қоғам өмірінде ақпарат ағымының таралуы табиғи тілде жүзеге асатындықтан, тіл білімінің қызметі күннен-күнге артуда. Тілдің жаңа компьютерлік формасы пайда болып жатыр. Осы жолдағы істің бірі қазақ тілі деректерінің терминологиялық банкін қалыптастыру екендігі сөзсіз.

Қазақ тілі деректерінің терминологиялық банкі (ДТБ) – бұл күрделі, ауқымды, ұдайы өзгеріп отыратын, көпаспектілі іздестіру мүмкіндігін

қамтамасыз ететін және жекелеген банктердің деректерін қолдану мүмкіндігі бар бірыңғай қызмет көрсету жүйесі. ДТБ әртүрлі салаларда қолданылатын терминдердің өзара салғастырылуын жеңілдете отырып, бүкіл қазақ терминологиясының бірізді сипатталуына қызмет етеді. Ол термин туралы көптеген деректерді жинақтаумен бірге, сол арқылы қазақ терминологиясының біркелкі сипатталуын мақсат етеді, тұтастай алғанда терминологиялық қызмет көрсету деңгейін арттырады. ДТБ өз жадына қазақ терминологиясының жаңа құрылымын енгізіп, әртүрлі принциптермен құрылған жекелеген банктердің дербес қызмет етуі деңгейінде анықтала қоймайтын, айтылымы жағынан да, мазмұны жағынан да сан алуан салалық терминологияның кірігуі нәтижесінде қалыптасатын жаңа құрылымдық байланыстарды айқындайды.

Қазақ тілінің кейбір мәселелерінің шешімін табуға ДТБ ықпал ететін бірқатар таза ғылыми міндеттер мыналар:

а) қазақ тілінің терминологиялық жүйесін ішкіжүйелер жүйесі ретінде модельдеу;

ә) жалпығылыми және жалпытехникалық тезаурустарды, автоматтандырылған білім банкін қазақ тілінің когнитивтік үлгілері ретінде құру;

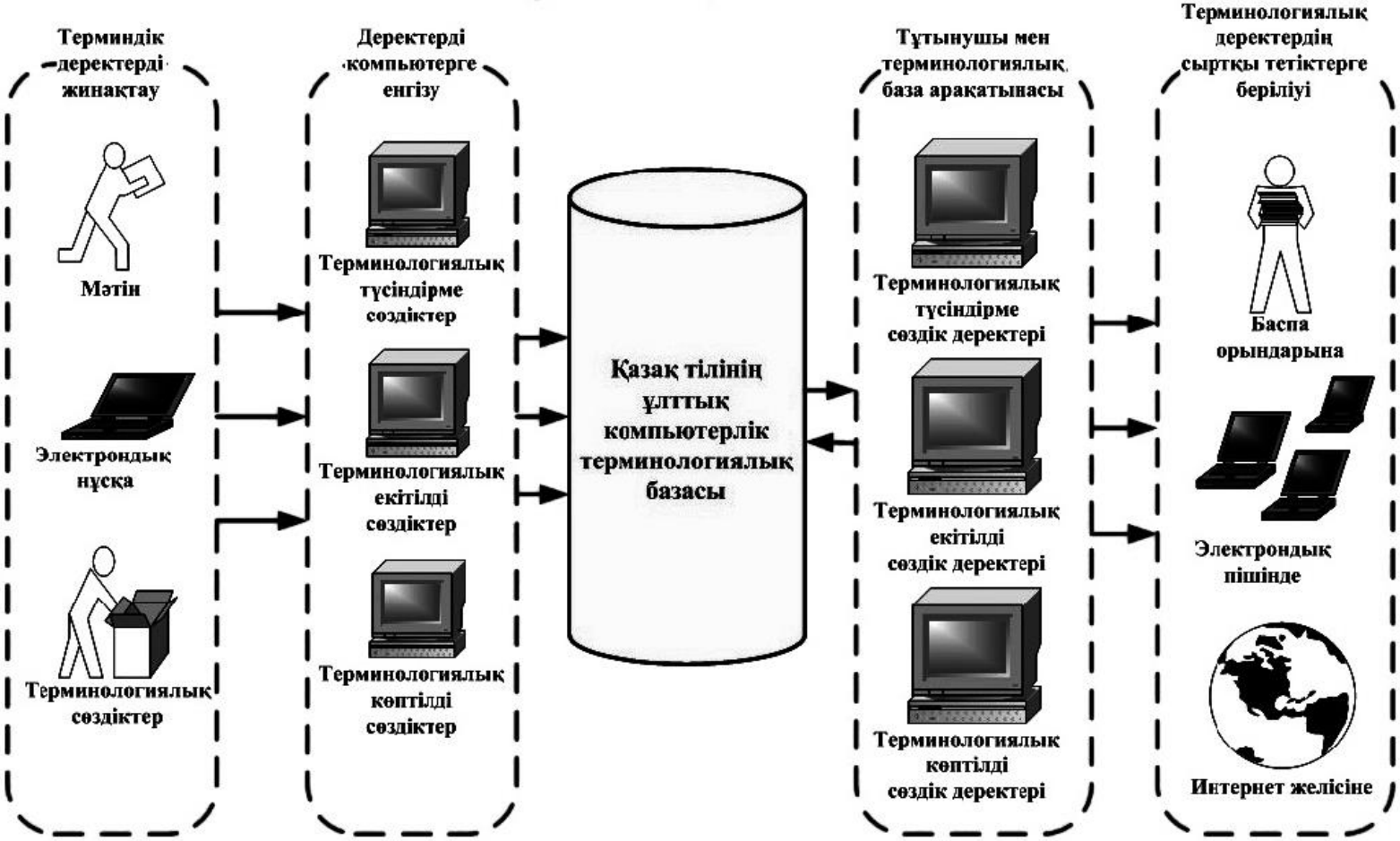
б) өзге тілдер терминдерінің аясында қазақ терминологиясын салғастыр-малы, типологиялық зерттеу, оның ішінде қазақ тілін мемлекеттік тіл ретінде зерттеу шеңберінде қарастыру.

Қолданыста бар анағұрлым танымал және жаңадан жобаланып отырған деректердің терминологиялық банктері жөніндегі әдебиеттерді талдау негізінде қазақ тілінің ДТБ-сын төмендегідей типтерге бөлуге болады:

- әртүрлі білім саласындағы мамандарға, сондай-ақ шет тілдерін үйренушілерге, студенттерге, әдеби редакторлар мен басқа да пайдаланушыларға анықтамалық-ақпараттық қызмет көрсету;
- ғылыми-техникалық әдебиетті дәстүрлі аударуды қамтамасыз ету;
- қазақ мәтінін өңдеудің автоматтандырылған жүйесін, ғылыми-техникалық әдебиетті компьютер арқылы аудару жүйесін, қазақ тіліндегі ақпараттық құжаттарды автоматты индекстеу мен автоматты рубрикалау, ақпараттық құжаттарды рефераттау жүйесін қамтамасыз ету;
- қазақ тіліндегі автоматтандырылған ақпаратты басқару, жобалау жүйелерін лингвистикалық мәліметтермен қамтамасыз ету;
- қазақ тіліндегі терминология жүйесін реттеу мен терминологиялық сараптамалар жасау жұмыстарын қамтамасыз ету;
- қазақ тілінің терминологиялық сөздіктерін түзу және оларды баспадан шығару процесін автоматтандыру;
- қазақ тіліндегі терминдерге қатысты анықтамаларды біріздендіру.

Қолданыстағы терминологиялық деректер банкінің көпшілігі ғылыми-техникалық әдебиетті аудару үшін қолданылатыны мәлім, алайда қазіргі кезде оған басқа да міндеттер жүктелуде, яғни ондай ірі деректер банкін пайдалануда көпфункционалдық үрдіс байқалады.

Ұлттық терминологиялық «Тіл-қазына» базасының «адам-компьютер» жүйесінде әрекет ету сызбасы



Қай тіл болмасын, оның деректерінің терминологиялық банкінің негізін құрайтын автоматты сөздіктер (АС) екені компьютерлік лингвистика саласы мамандарына бұрыннан да мәлім [1]. Сондықтан оқырман қауымына ұсынып отырған еңбегімізде қазақ тілі деректерінің терминологиялық банкін құрастырудың алғышарты ретінде автоматты сөздіктер (АС) мен ДТБ мәселесінің кейбір теориялық, практикалық жақтарына және олардың қызмет ету аясына тоқтала кетуді жөн көрдік. Автоматты сөздіктерді (АС) негізгі үш типке бөліп қарастыруға болады. Міне, осы тұрғыдан алғанда АС-ның бірінші типіне: жиілік сөздіктер, сөзнұсқағыштар және конкорданстарды жатқызуға болады. Қазіргі кезде мұндай сөзтізбелер түрлері біртұтас автоматтық сөздікті құрайды. Ал оның сөзтізбесі сөзқолданыстың жиілігімен немесе басқа да белгілермен жабдықталады, сонымен бірге олар деректердің мәтіндер сақталатын базасына ену әрекетіне қажеттілік ететін индекстер рөлін де атқара алады.

Сөздіктердің екінші типі деректердің терминологияылық банкі (ДТБ) түрінде көрініс табады. Ал оның үлкен көлемдегі бірнеше түрін атауға болады:

- ДТБ-ның үлкен көлемді 8 тілді және 1,5 млн. терминдерді қамтитын TEAM банкі;
- 7 тілді және 1,5 млн. терминдерді қамтитын LEXIS банкі;
- 300 мың сөздік мақаланы қамтитын EURODICAUTOM (Люксембург) атты көптілді терминологиялық банкі және т.б.

Бұл аталғандардан басқа Франция, Швеция, Канада, Германия және т.б. елдердегі аз көлемдегі саны 40-тан асатын терминологиялық деректер банктері жүзеге асырылғаны белгілі. Соңғы жылдары мұндай ТДБ-ның саны мен сапасы және олардың сан алуан атаулары да күннен-күнге өсіп келеді.

ДТБ-лардың елеулі ерекшелігі ретінде сөздік мақаланың белгіленген (тұрақты) пішімде, яғни оның сөз-термин немесе терминдік тіркестерге тән «досьесінің» (құжаттар жинағының) болуын айтуға болады. Дәлірек айтқанда, бұл ғылыми-техникалық әдебиеттерді бір тілден екінші тілге аударуға қажетті аудармашыларға қызмет ететін құжаттық ақпарат-іздестіру жүйесі. Аталған жүйеде реферат түріндегі құжаттарда терминдерге мынадай сипаттама беріледі: қызмет көрсету белгілері (сөздік-дереккөздің коды, терминдік жүйенің коды, автордың аты-жөні мен жауапты мекеменің аты және т.б.); қарастырылып отырған лексика-семантикалық варианттың лингвистикалық сипаттамасы (сөздің формасы, дифиниция, мысалдар, синонимдер және т.б.); осындай сипаттамалардың басқа тілдерге қатысты баламалары. Егер қайсыбір сөз-терминнің сол тілде бірнеше мағынасы болатын болса, онда барлық мағыналар соншалықты «досьеде» тіркелуі қажет.

Автоматты сөздіктердің үшінші типі деректер базасында орналасқан жалпытілдік моно- немесе көптілдік сөздіктер. Мұндай базалардың елеулі ерекшелігі – сөздік мақаланың белгілі сегменттерде сақталу мүмкіндігі. Нәтижесінде, сөздік мақаланы шартты түрдегі еркін форматта және сөздік мақаланы белгіленген құрылымдық пішімде сақтауға мүмкіндік туады.

Барлық автоматты сөздіктер көлем жағынан да (әрбіреуі 60 мыңға жуық сөздік мақаладан тұрады) және атқаратын қызметі бойынша да шамамен бір-біріне ұқсас болып келеді. Мысалы, атқаратын қызметіне қарай: сөздік мақалаға қатысты жеке және топтық сұраныстарға жауап беру (жауапты экранға немесе қағаз бетіне шығару), әліпбилі немесе мәтінге бағытталған сөздікшелерді құрастыру, сөздік мақалаға жекебасты түзетулер жасау, сөздікті жеке немесе топтық деңгейде толықтырулар жүргізу.

Деректердің терминологиялық банкінің (ДТБ) көпшілікке арналған автоматты сөздіктерден (АС-дан) негізгі айырмашылығы – оларды пайдалану жағдаятында. Мәселен, егер терминологиялық банк негізінен аудармашы қауымға қызмет ету бағытын ұстанса, екіншілер, яғни автоматты сөздіктер, оларды құрастыру мен қолдану кезінде автоматтанған әрекет ету бағытын ұстанады. Көпшілікке арналған автоматты сөздіктерде «айналымдылық» мәселесі жеңіл шешім табатыны белгілі. Мысалы, орысша-қазақша сөздікті қазақша-орысша сөздікке (немесе керісінше) «айналдыру» әрекеті автоматты сөздіктерде қиындық тудырмайды. Сонымен бірге көпшілікке арналған АС-ны әртүрлі сервистік қызметі мен сөздік мақала құрылымының кеңейтілген жағдайында деректердің терминологиялық базасы (ДТБ) ретінде де пайдалануға мүмкіндік бар. Ал ДТБ-ны өз тарапынан лексикографиялық жұмыстарды автоматтандыру жүйесі, яғни қайсыбір лексикалық бірлік жайлы мәліметтерді жинақтайтын құжаттардың ақпараттық жүйесі ретінде (сөзге «досье» ретінде) пайдалануға болады [2].

Сондықтан картотекалық қорды автоматтандыру жүйелерін (КАЖ), автоматты сөздіктер (АС) мен деректердің терминологиялық банкін (ДТБ) лексикографиялық жұмыстарды автоматтандыру жүйесі немесе автоматтанған лексикографиялық жүйе (АЛЖ) деп қарастыруға да болады.

Автоматтанған лексикографиялық жүйе (АЛЖ) төмендегідей лексикографиялық амалдарды орындай алады:

- сөзтізбені іріктеу (кейбір сандық көрсеткіштер, мысалы, қолдану жиілігі, автоматты сөздіктердегі лексеманы іздеу саны және т.б.);
- бір топ екітілді сөздіктер үшін біркелкі сөзтізбе қалыптастыру;
- сөзтізбені автоматты түрде морфологиялық ақпаратпен қамтамасыз ету;
- иллюстративтік материалды таңдау (пәндік және терминологиялық лексикаға қатысты беделді дереккөздердегі дефинициялар);
- сөздіктің түрлі белгілеріне қарай, мысалы, сөздік мақаланың құрылымы мен анықтамалардың логикалық тұрпатын салыстыра келе, сөз топтары бойынша сөздікшелер құрастыру;
- алдын ала белгіленген мәтіндерді топтастыруға қатысты стилистикалық таңбалардың сәйкестігін бақылау;
- сілтемелерді бақылау, жетекші және сілтемелік сөздік мақалаларды салыстыру;
- сөздік жұмысына ақпараттық-анықтағыштық тұрғыда қолдау көрсету (лексикалық бірліктердің кейбір аспектілері сипатталған әдеби дереккөздерді іздестіру және т.б.).

Аса жетілген автоматты сөздіктерді лексикографиялық дереккөздің жаңа типі деп те атауға болады. Себебі автоматты сөздіктерді сөздіктердің жеке бір түрі деп ұғынудан гөрі, белгілі мақсат үшін пайдалану сөздігі деп түсінген жөн сияқты.

Тұтынушы-лексикограф автоматты сөздікте орын алатын мақала ақпаратын әртүрлі көзқараста ұғынуы мүмкін, мысалы:

- егер деректер базасындағы тиісті грамматикалық (фонетикалық) ақпаратпен жабдықталған сөздік мақаланың тақырыбына ғана назар аударатын болсақ, ол сөздікті орфографиялық, немесе орфоэпиялық, немесе фонетикалық, немесе грамматикалық сөздік деп ұғынуға болады.

- егер жекелік сауалға қатысты сілтеме түріндегі сөздік мақалалар тобы сұраққа жауап түрінде берілсе, автоматты сөздікті синонимдер (антонимдер), омонимдер, сөзжасамдық немесе идеографиялық сөздік деп те ұғынуға болады;

- егер сөзге қатысты сауалда тұтынушыны сөздік мақаланың қай бөлігі қызықтыратынына байланысты, оны дифинициалық немесе аударма сөздік деп қарастыруға болады;

- егер қайсыбір сөзге қатысты қойылған сауалға жауап ретінде барлық фразеологиялық тіркестер тұтынушы назарына ұсынылатын болса, онда автоматты сөздікті фразеологиялық сөздік деп ұғынуға болады;

- егер қайсыбір сөзге байланысты ақпарат сөздік базасының құжаттық-ақпараттық бөлігіне сәйкес келіп, оған қатысты барлық «досье» берілетін болса, онда оны тарихи немесе этимологиялық сөздік деп те қабылдауға болады және т.б.

Автоматты сөздіктер жүйесіндегі тілдік процессорларға қатысты мынадай амалдар орындалады: жалғауларды диагностикалық тәсіл бойынша автоматты түрде қалыптастыру; тұрақты сөз тіркестерін туындату; тіркесім мен меңгеру үлгісін айқындау, берілген құрылымға байланысты қолдану үлгісін таңдау және т.б.

Автоматты сөздіктердің басым көпшілігінде типтік айырым-белгі жасырын тұрады, сол себепті сөздіктер типологиясына қатысты мәселені есептеуіш лексикография саласы төмендегідей түсіндіреді:

1. Сөздік мақаланың орналасу тәртібімен сипатталатын барлық сөздіктер деректер базасында басқаша пішімде орналасса да, тұтынушы-лексикограф сөздіктерді өзінің қалауынша, яғни тура әліпбилі және кері әліпбилі, ұялы, тақырыптық және дескрипторлық, әліпбилі-жиілікті және жиілікті-әліпбилі тәртіптері бойынша автоматты сөздік базасынан шығарып алуына мүмкіндігі бар. Ал аталған сөздік түрлерін сөздіктің сыртқы тұрпаты параметрлеріне жатқызады.

2. «Біржақты» деп аталатын барлық сөздіктерді (орфографиялық, фонетикалық, орфоэпикалық, грамматикалық, кері әліпбилі сөздік) толық түрдегі «екіжақты» (мысалы, екітілді аударма немесе түсіндірме) сөздіктер базасынан автоматты түрде құрастырып, тұтынушы назарына ұсынуға болады.

3. Сапалы құрастырылған түсіндірме сөздіктер негізінде сөзжасамдық

пен фразеологиялық, синонимдер, антонимдер, омонимдер, неологизмдер мен архаизмдер сөздіктерін де автоматты түрде құрастыруға мүмкіндік бар.

4. Деректердің сөздік базасында орналасқан сапалы энциклопедиялық сөздіктер негізінде барлық қысқартуларды, жалқы есімдер мен номенклатуралық анықтағыштарды автоматты түрде құрастыруға да болады.

Сонымен, қазақ сөзінің «Тіл – қазына» деректер базасының терминологиялық банкінің мақсаты «адам-компьютер-адам» жүйесіндегі ғылыми-техникалық сөйлеу актілерінде тілдік норманы жүзеге асыру болып табылады. Алайда тілдік норманы жүзеге асыру ғылыми-техникалық тілдің барлық элементтерін тізбелеу негізінде осы норманы алдын ала анықтап алмайынша мүмкін болмайды.

Қазақ сөзінің «Тіл – қазына» деректер базасының терминологиялық банкінің тізбелеу қызметі мынаған бағытталған:

- терминологиялық лексиканы барынша толық қамту және жан-жақты сипаттау;
- терминологиялық лексиканың барлық қолданылу аспектілерін көрсететін алуан түрлі дереккөздерді барынша енгізу;
- терминологиялық жүйелерді құрастыру мен дамытуға ынталы барлық ұйымдармен қарым-қатынас жасау.

Қазақ сөзінің «Тіл – қазына» деректер базасының терминологиялық банкінің тізбелеу қызметі тек ғылыми ынталылықпен ғана емес, сонымен қатар ақпараттық орталықтарды, баспаларды және басқа да ұйымдарды өздері жасаған тілдік құралдарды ұсынуға міндеттейтін ұйымдастырушылық шешімдермен де нығайтылуы қажет.

Қазіргі жаһандану кезеңіндегі Тәуелсіз Қазақстан Республикасының ғылыми-техникалық жетістіктеріне, экономикалық өркендеу қарқынына сәйкес, «адам-компьютер-адам» қарым-қатынас саласындағы тілдің қолданылуын анықтай отырып, қазақ сөзінің «Тіл – қазына» деректер базасының терминологиялық банкі ғылымның барлық салаларына жаппай ықпал ете алады деген сенімдеміз.

«Тіл – қазына» деректер базасын құру – терминологиялық қызмет жасаудың жаңа бағыты. XX ғасырда басталған ғылыми-техникалық «революция» әлемнің кез келген мемлекетінің ішкі-сыртқы саясатына, әсіресе экономикалық әлеуетіне ерекше серпін бергені белгілі. Бұл Тәуелсіз Қазақстан Республикасы сияқты дамушы елдердің жас мемлекет ретінде қалыптасуында айрықша рөл атқарды. Қоғамдық қызметтің қай саласында да қолданбалы бағыт басымдық алды. Осы орайда лингвистиканың қолданбалы саласы да қалыптасып, дәстүрлі тіл білімінің бағыттарын өз әдіс-тәсілдері зерттеуге кірісті.

Қолданбалы лингвистиканың мәселелерін зерттеуде оның терминтану жағына айрықша көңіл бөлінуі қажет. Терминтану қазіргі заман сұранысына сай ғылыми пән және оның зерттеу нысаны терминдер мен терминологиялық жүйелер. Аталған ғылым саласы XX ғасырдың ортасында тіл білімінің, логиканың, информатиканың, психологияның және басқа да ғылымдардың түйісу жерінен пайда болды. Қай тілді алсаңыз да терминдер – сөз немесе сөз

тіркестер түрінде кездеседі де осы тєріздес басқа сөздердің заңдылығымен бірдей қолданыс табады. Сондықтан терминтану пәнінің әдіс-тәсілдері тіл білімі ұстаным ететін зерттеу жолдарымен ұқсас болып келеді. Мысалы, терминтану пәні де лингвистикадағыдай теориялық жєне қолданбалы деп аталатын екі бөлімнен тұрады.

Терминтанудың қолданбалы жағын сөз еткенде, ең алдымен, оның терминтану саласындағы мәртебесін анықтап алу қажет. Тіл біліміндегі қолданбалы лингвистиканың өз теориясы болатындығы сияқты қолданбалы терминтанудың да теориясы жєне оның практикалық терминологиялық қызметінің бірқатар бағыттары бар. Бұл бағыттардың өздеріне тән терминологиялық мектебі бар десек те болады. Сондықтан терминологиялық бағытты таңдау оның қолдану қағидаттарына, әдістеріне, көлеміне, қызмет ету деңгейіне, халық шаруашылығындағы нәтижелеріне, экономикалық сипатына және т.б. байланысты болады. Осыған орай, терминтану мектептерінің практикалық қызметінің мынадай түрлерін атауға болады [3]:

- лексикографиялық терминология қызметі;
- термин мен терминдік жүйелерді жетілдіру (реттеу, тұрақтандыру, редакциялау);
- ғылыми-техникалық және басқа да маманданған мәтіндер аясындағы терминдерді аудару;
- терминологиялық деректер банкін құру (ТДБ);
- терминологиялық ұйымдар мен орталықтардың методологиялық (ғылыми) қызметін ұйымдастыру және т.б.

Практикалық терминологиялық қызмет өзара тәуелді екі түрлі аспектіге ие: оптимизациялық және нормативтік. Бір жағынан алғанда, ондай қызмет терминдер мен олардың мағынасын дәл анықтайтын және тұтынушыны қанағаттандыратын тиімді вариантты таңдап алуға бағытталған. Екіншіден, таңдалып алынған тиімді терминдер мен олардың анықтамалары нормативті құжаттар мен бұйрық-жарлықтар қағаздары арқылы бекітіліп, баянды етіліп отырады. Оларға жататындар: стандарттар, терминологиялық кепілдемелер, нормативті сөздіктер, редакторлық нұсқаулар мен ережелер және т.б. Бұл құжаттар мен бұйрық-жарлықтар терминологиядан тыс жатқан лексика аясында қолданыс табатын нормаға қарағанда әсерлілеу, күштілеу деуге болады.

Соңғы кезде ғылыми-көпшілікке арналған терминологиялық сөздіктер де жарық көруде. Бұл сөздіктерде әр ғылым саласының терминдік жүйесі көрініс табады. Әдетте мұндай сөздіктер әліпби тєртібімен емес, жүйелік ұстанымға негізделіп құрастырылады.

Мақсаты жағынан мынадай терминологиялық сөздіктерді атап өтуге болады: «узус» сөздіктері (тіркеу, түгелдеу сөздіктері), регламенттеу, оқу жұмысы, хабарлау, жүйелеу мақсатындағы сөздіктер.

Ғылыми-техникалық лексикография негіздерін толығырақ меңгеру үшін А.С. Гердтің «Основы научно-технической лексикографии: как работать над терминологическим словарем» атты еңбегіне сілтеме жасағанды жөн көреміз [4].

Термин мен терминдер жүйесін бір ізге келтіру. Бұл бағыт қолданбалы терминтану саласының негіздерінің бірі болып саналады. Бірізге салу жұмысының жемісі ретінде регламенттеу сөздіктерін, нормативті және нормативті емес сөздіктерін атауға болады. Нормаға қатысты сөздіктерде терминдердің әр категорияларына стандарт көрсетіледі. Сонымен бірге бұл реттегі орындалған жұмыстарға ұсынылған терминдердің жинағын да жатқызуға болады.

Реттелген терминдер ұсынылған (бекітілген) терминдер жинағы түрінде жарық көреді. Қазақстанда мұндай жұмыстарды Мемлекеттік терминология комиссиясы атқарады. Бұл бағытқа, яғни терминдер мен олардың жүйесін бірізге келтіру мен оларды арнайы құжаттар арқылы бекітуге қарсы пікірлер де жоқ емес. Бірақ машина мен механизмдердің бөлшектері мен тораптарын стандарттау қаншалықты қажет болса, терминдерді де бірізге келтіру ғылым мен техниканың алға басуына соншалықты қажет деуге болады. Егер терминдердің атаулары арасындағы синонимдік сипат жойылатын болса, онда ол ғылыми қызметтің «тоқырап қалуына» әкеліп соғады деген күдік те туындауы мүмкін. Ал бірізге түскен терминдерге қатысты нормативті құжаттардың уақыт пен кеңістік бойынша жєне оның қызметіне қарай шек қойылу жағдаяты жоғарыда аталған күдікті жоққа шығара алады.

Терминдерді бірізге салудың тағы бір оң қызметі – жүйелілікке келтіру мүмкіндігі. Стандарттар мен ұсынылған терминдердің жинағында, сөздіктерге қарағанда аталған қызмет басым байқалады. Стандарттау мен ұсынылған терминдерді жинақтауды дұрыс жолға салу нәтижесінде әрбір терминдік жүйедегі орын бір ғана лексикалық бірлікпен толтырылып, оған сәйкестікке бір ғана анықтама беріледі (сөздіктің оң жағындағы мақаласы). Әрине, бұл әрекет, белгілі білім аясына жататын барлық таңбалау құралын жүйеге келтіреді және мұндай білім аяларына шектеулер қойып, олардың өзара байланыстығы анықталады. Терминдерді бірізге салуда мұндай шараның да маңызы зор.

Соңғы жылдары терминдер мен терминдік жүйені бірізге салудың үшінші түрі орын ала бастады, ол – ұлттық және халықаралық деңгейдегі үйлестіруді іске асыру. Әрине бұл әрекет бойынша әртүрлі тілдердегі және жақын білімдер аясындағы терминдерді бүтіндей теңестіру мақсаты қойылмайды. Дегенмен, терминжасау мамандары мен әр ғылым салаларының өкілдері әр дәрежеде терминдердің өзара үйлесімділігін іздестіруге және оның шегін анықтауға атсалысады. Терминдер қолданыс табатын тілдердің өзара ерекшеліктері ескеріледі – сөзжасам тәсілі, фонетикалық айырмашылы, сөздердің тіркесім заңдылықтары және т.б. Сонымен бірге әрбір терминдік жүйе белгілі теорияға, тұжырымдамаға, көзқарастар жүйесіне тәуелді екені де ескеріледі.

Терминдік жүйелердің халықаралық үйлесімділігінің экстралингвистикалық алғышартына жататындар: білімнің интеграциясы, ғылым мен технологияны интернационалдау, ғылым мен технологияның теориясы мен әдістемелік базасының бірлігі. Терминдік жүйелердің

халықаралық үйлесімділігіне мүмкіндік туғызатын лингвистикалық факторларға жататындар: арнайы мақсатпен бір-біріне жақын жатқан табиғи тілдер үшін ортақ тіл туындату, арнайы мақсаттағы тілдердің лексикалық бірлігінің функциясы ретінде орын алатын терминдерді интернационализациялаудың жедел әрекеттері.

Терминдік жүйелердің халықаралық үйлесімділігінің негізгі екі қағидаты бар:

а) егер екі не одан да көп мемлекеттерде білімнің белгілі бір саласы немесе сол саладағы зерттеулер бірдей дәрежеде дамыған болса, аталған үйлесімділікті қабылдаудың мүмкіндігі бар жєне ол мақсатқа сай келеді;

ә) егер бұл мемлекеттерде белгілі бір білім саласының негізінде бірдей не жақын жатқан ұғымдар жүйесінің ұқсас (бірдей) немесе жақын жатқан теориясы болса, аталған үйлесілімділікті қабылдауға мүмкіндік бар.

Аталған ұстанымдар, белгілі бір дәрежеде, іске асып жатса, терминдік жүйелердің мынадай үш дәрежедегі үйлесімділігіне қол жеткізуге болады:

1) теориялардың ішінара сәйкес келуі мен тілдік құралдардың айырым табу жағдайындағы ұғымдар жүйесінің ортақтығы. Бұл жағдайда, үйлесімділік тек ұғымдар жүйесі мен оларды белгілейтін терминдерді өзара байланыста ұстайтын тұрпатта түсінік табады;

2) теориялар мен тілдік құралдардың айырым табу жағдайының толық сәйкес келуіндегі ұғымдар жүйесінің ортақтығы. Бұл жағдайда, үйлесімділік тек ұғымдар жүйесінің бірдей келуі мен оларды белгілейтін терминдердің құрылымдық-семантикалық жақындығы түрінде түсіндіріледі;

3) теориялар мен тілдік құралдардың толық сәйкес келу жағдайындағы ұғымдар жүйесінің ортақтығы. Бұл жағдай, үйлесімділік ұғымдар жүйесі мен терминдерді интернационализациялаудағы мазмұнның көлемі мен өзара байланысының бірдейлік сипатымен көрініс табады.

Ғылыми-техникалық және басқа да арнайы терминдерді аудару мәселесі. Арнайы мәтіндерді аударуда, оларда қолданылған терминдерді барабар аудару ең маңызды іс болып саналады.

Аудармаға арналған көптеген ғылыми конференциялар мен семинарларда терминдерді аудару тәсілін таңдау проблемасы төңірегінде өткен пікірталастардың маңызы зор. Қазіргі кезде ғалымдардың осындай әртүрлі тәсілдерді қолдану жайлы ұсыныстары зерттелуде [5]. Оларды қысқаша былай тұжырымдауға болады:

а) Ең алдымен аударма тілі (АТ) және түпнұсқа тілі (ТТ) деген атаулардың мағынасын ажырата білуіміз керек. Егер «аударма тілінен» (АТ) белгілі бір тәсілге сүйеніп, «түпнұсқа тіліне» (ТТ) сәйкес келетін қолайлы балама табылып жатса, сол тәсілді оңтайлы тәсіл деп атауға болады. Мұндай тәсілді қолдану үшін тіл иелеріне қатысты мемлекеттердің қоғамдық даму деңгейлері бірдей немесе өзінің тарихи кезеңдерінің бірінде одан өтіп кеткен болуы керек. Мысалы, екі тілдегі балама-терминдердің ерекшелігі мынада. Әдетте, әр тілдегі балама-терминдер екі тілде де бұрын қолданылған болса, аударма жасау кезінде жаңа терминдер мен неологизмдер сияқты оларды қайта жасамай, баламаларын қайтадан іздеп-табуға әрекет етуіміз керек.

ә) Аударма тіліндегі (АТ) жаңа терминді жасау үшін осы тілде бар сөз не сөз тіркеске түпнұсқа тілі (ТТ) әсерінің арқасында жаңа мағына берілуі қажет. Мысалы, орыс тіліндегі ботаникалық термин *дерево* жаңа мағынаға ие болып, жаңа ғылыми салалардың – топтастыру теориясы, басқару теориясы, ақпарат теориясы терминдеріне айналып отыр. Оны мынадай терминдік сөз тіркестерден байқауға болады: *дерево целей, поисковое дерево* (ағылшын тілінің *tree* сөзінің әсерінен) және т.б. Орыс тілінің саяси термині *«разрядка»* бұрын физика саласында жиі қолданылып жүрген терминнің жаңа мағынасы деуге болады. Бұл сөздің жаңа терминдік мағынаға ие болуы француз тілінің саяси мағынадағы *detente* сөзіне қатысты екен. Себебі *detendre* – «*ославлять, разряжать*» етістігінен жасалғаны анық.

б) Егер аударуға тиісті лексикалық бірлік екі тілде де сәйкес келіп жатса, оларды ***семантикалық калька*** деп атайды. Семантикалық кальканы пайдаланғанда түпнұсқа тіліндегі (ТТ) терминнің құрылымы түпнұсқа тілінің нормасына сәйкес келеді, ал аударма тілінің (АТ) жасалатын термин құрылымы – аударма тілінің (АТ) нормасына сєйкес келеді. Техникалық ғылымдар саласында құрылым жағынан күрделі болып келетін терминдердің әр элементін бөлек-бөлек аудару (калькілеу) тәсілі кең қолданылады. Мысалы, ағылшынша: *Swithing diagram – коммутационная схема, motor selector – маторный искатель* (электроника бойынша ұсынылған мемлекеттік стандарттан алынады).

в) Егер аударма кезінде лексикалық бірлік сол құрылымы қалпында басқа тілдегідей қайталанып пайдаланылса, ол құрылымдық калька немесе кальканың өзі деп саналады. Кейбір басқа тілден ауысып алынған құрылымдық модельдер калькілеу нәтижесінде аударма тілінде (АТ) тұрақталып қалады, ал басқалары жат қалпында қалып, жекеленген құрылымдық калька ретінде қолданыс табады. Міне, калькілеу тәсілінің ұтатын да, ұтылатын да жері осында.

г) Егер аударма кезінде түпнұсқа тіліндегі (ТТ) терминдердің әрі семантикасы, әрі құрылымы, әрі тұрпаты (дыбыстық құрамы мен жазылуы) толығымен аударма тіліне (АТ) өзгеріссіз ауысса, ол кірме терминдер болып саналады. Осындай жағдайда, кірме терминдерді екі тілдің тікелей түйісуінен болатын және интернационалдық құбылыстардан нақты ажырата білген жөн. Ең бірінші, кірме терминдерді классикалық жасалу тәртібіне негізделген еуропа мәдениетінің ерекшелігімен анықталатын грек-латындық элементтер арқылы құрылған тілдік бірліктермен шатастырмауымыз керек.

д) Көптеген жағдайда термин сөз – сипаттау құрылымымен бірге аударылғаны дұрыс деп саналады. Бұл аудару әдісі, ең алдымен, нақты мемлекеттің ақиқат өмірін суреттейтін эквивалентсіз терминдер үшін қолданылады..

Терминологиялық редакциялау. Мәтіндерді әдеби жєне техникалық редакциялаумен қатар терминологиялық редакциялау – мәтінді ғылыми редакциялаудың бір түрі. Терминологиялық редакциялауды қарастыру үшін оның мақсаты мен міндетін, негізін және әдістерін сипаттап алу қажет.

Редакциялаудың басқа да түрлері сияқты оның мақсаты – мәтіннен танымдық, білім беретін немесе тәрбие алатын, яғни барынша толық түрдегі күшті әсер алатын зор нәтижеге жету.

Терминологиялық редакциялау барысында алға қойған мақсатқа жету үшін ерекше материал ретіндегі терминдермен әрекет етуге тура келеді. Ал терминдер, жоғарыда айтылғандай, белгілі тілдің табиғатына, оның терминдік жүйеден алатын орнына көп байланысты. Сондықтан терминологиялық редакциялаудың міндеті – ол, терминдер мен сөздерді және олардың орнына жүретін тілдік бірліктерді барынша орынды пайдалануға мүмкіндік туғызу. Терминдерді таңдауда, оның нақты мәтіндегі ақпараттық рөлін және сол ғылым не техника саласындағы басқа терминдермен салыстырғандағы орнын ескеру қажет болады.

Терминологиялық редакциялаудың қағидаттары көптеген жағдайда шығарманың стилі мен жанрына байланысты болып келеді. Мысалы, ғылыми мәтіндерде (ғылыми мақала, монография) арнайы терминдер көптеп кездеседі де, бірақ маман-оқырман оларды жақсы білуі керек деген көзқараспен терминдерге арнайы түсініктеме берілмейді. Ал ғылыми-көпшілікке арналған мақалалар мен газет бетіндегі очерктерде терминдерге тиісті сөз орамдарымен сипаттамалар беріліп отырады. Себебі көпшілік оқырман мәтінде баяндалған ғылыми проблемамен, саяси немесе экономикалық оқиғалармен бірінші рет танысып отыруы да мүмкін.

Терминологиялық редакциялаудың әдістеріне келетін болсақ, олардың саны көп жағдайда екіден аспайды. Біріншісі бойынша, редактор мәтінде қолданыс тапқан терминдерді тиісті анықтамалармен немесе бекітілген нормативті құжаттармен (стандарттар, ұсынылған терминдер жинағы), не болмаса осы ғылым мен техника аясында кең қолданылатын терминдермен сәйкестікке қояды. Ал, екінші әдіс бойынша, орын алатын әрекет мынаған тіреледі: егер редакцияланудан өтіп жатқан мәтін ішінде жаңа терминдер кездессе, автор олардың не себепті орын алғанын түсіндірумен қатар, әрбіреуіне анықтамалар беріп отырады.

Қазақ тіліндегі ғылыми-техникалық стиль және оған тән терминдік жүйе жаңадан ғана қалыптасып келеді деуге болады. Себебі техникалық ғылымдардың қазақ тіліндегі көрінісі немесе дерек көздері түрлі монографиялық зерттеулер, оқулықтар, мақалалар, баяндамалар, тезистер, патенттер т.б. сан-салалы еңбектер арқылы танылады. Ал бұл аталғандардың саны мен сапасы қазақ ғалымдарының әлі де болса арнайы көңіл бөлуін қажет етеді. Ана тілімізде жарық көре бастаған жаңа бастаманың нәтижесінде тілімізге жаңа ұғымдар мен техникалық жаңа атаулар келе бастағаны белгілі. Сондықтан қазақ тілінің жалпы лексикография саласының жаңа бағыты ретінде ғылыми-техникалық лексикография саласы да өз отауын құра бастады деуге болады.

Кез келген тілдің дәстүрлі түсіндірме сөздіктеріне қарағанда ғылыми-техникалық терминологиялық сөздіктер, ең алдымен, белгілі білім аясы мен техникалық саланың мамандарына арналатыны белгілі. Терминологиялық сөздіктер, өздерінің мақсатына байланысты, жоғары білімді мамандар мен

ғылыми әдебиеттерді аударумен айналысатын кәсіптік аудармашыларға арналып құрастырылады. Жоғары дәрежелі мамандарға мұндай сөздіктер тек қажетті деген сөзді тауып алуға ғана мүмкіндік туғызып қоймай, сонымен бірге оның ұғымын анықтауға және әртүрлі басылымдарда кездесетін кейбір синонимдік сөздер мағыналарын ажыратуға да мүмкіндік туғызады.

Ғылыми мақалаларды бір тілден екінші тілге аудару кезінде идеографикалық принципте құрылған біртілді және көптілді түсіндірмелі-аударма терминологиялық сөздіктер де аса қажет [6; 7].

Ғылыми-техникалық терминологиялық сөздіктерді орта және жоғары техникалық оқу орындарының студенттеріне, әр салалық өндіріс орындары мен ауылшаруашылық аясындағы сан-салалы мамандарына және радио мен баспа қызметкерлеріне және т.б. арналған сөздіктер деп те ажыратуға болады. Сол сияқты, мектеп оқушыларына арналған терминологиялық сөздіктер де жеке пішінде шығарылып жүр. Бұл сөздіктердегі деректерді оқырманға ұсыну тәсілінде оқушының жас ерекшелігіне айрықша көңіл бөлу негізгі ұстаным болып табылады. Мысалы, терминологиялық сөздікте қызықтыру элементтерімен бірге суреттердің, сызбалардың орын алуы. Соңғы кездерде ғылыми-көпшіліктік терминологиялық сөздіктермен қатар бөлек ғылыми аяның терминдік жүйесін анықтайтын арнаулы ғылыми сөздіктер де жарық көре бастады. Әдетте, мұндай сөздіктердің құрылымы әліпби тәртібінен гөрі, жүйелілік принципке негізделіп жасалады.

Қолданылу аясына қарай, тіркейтін, түгелдейтін – «узус» сөздіктерін, белгілі ережелерге бағындыруды көздейтін – регламенттеуші сөздіктерін, оқу жұмысын көздейтін – мәлімет беру, жүйелендіруші терминологиялық сөздіктер де жарық көруде [8].

Стандарт терминдердің анықтамалары бойынша олар – ресми, мемлекеттік-құқықты статусқа ие болатын жүйелі типтегі терминологиялық сөздіктер. Терминологиялық стандарттардың құрамы – кіріспе, негізгі және әліпбилік көрсеткіштерінің бөліктерінен тұрады.

Кіріспе бөлікте стандарттың таратылу аясы, стандарттау нысанын анықтау және қолдануға болмайтын терминдер, синонимдер мен аударма баламалары жайлы мәліметтер беріледі.

Стандарттың негізгі бөлімінде осы стандартқа қатысты қабылданған ұғымдар жүйесіне сәйкес орналасқан терминдер мен олардың анықтамалары қамтылады.

Терминологиялық сөздіктерге әр білім саласы бойынша Ғылыми-техникалық терминологиялық комитетінің жүйелі түрде шығарып тұратын ұсынылған (бекітілген) терминдердің жинағын да жатқызуға болады. Мұндай жинақтарда негізгі терминдермен бірге олардың анықтамалары, синонимдері және аударма баламалары да беріледі.

Нормативті терминологиялық сөздіктерді құрастыру аса маңызды іс. Сондықтан ондай сөздік бір ғана терминдік жүйеге қатысты болуымен бірге олардағы терминдердің қолдану шарттылығы арнайы құжат арқылы бекітілуі қажет.

Арнаулы ғылым-техникалық терминологиялық сөздіктерді құрастырудың алға қойған мақсатына қарай, олар – **ақпараттаушы** және **жүйелендіруші** деп екіге ажыратылады.

Арнаулы ақпараттаушы сөздікке мысал ретінде ақпараттық-іздестірудің тезаурустық түрін атауға болады. Ақпараттық-іздестіру тезаурусы деп біртілді қадағаланып отыратын және лексикалық бірліктерге өзгеріс енгізіп тұру сипатындағы сөздікті атайды. Бұл сөздік бір ғана тілдің лексикасына негізделіп, ақпаратты өңдеу мен іздестіру қажеттігі үшін құрастырылады. Ақпараттық-іздестіру тезаурусы, ол – ақпаратты автоматты өңдеу мен іздестіруді жеңілдететін сөздік түрі. Ондай тезаурус терминдер арасындағы мағыналық, семантикалық қатынасты да толық ашуға жәрдем етеді. Ақпараттық-іздестіру тезаурустың негізгі бірлігі – терминдер мен дескрипторлар. Оның әліпбиленген лексика-семантикалық бөлігі – бір жерге жинақталған дескрипторлық сөздік мақалалары. Әрбір дескрипторлық сөздік мақала термин-дескриптордан бастама алады да, ал одан кейін дескриптордың мақала шеңберіндегі синонимдері беріледі және негізгі дескриптормен грамматикалық немесе ассоциативті қатынаста болатын басқа да терминдер орналасады [9].

Жүйелендіруші арнаулы сөздіктерге жататындар: *классификаторлар, рубрикаторлар, жиілікті сөздіктер* және *терминдік жүйе* сөздіктері. Классификаторларда реттелген түрде ғылым мен техниканың жүздеген нысан атаулары және олардың сипаттамалары беріледі. Мұнда ең алдымен техника-экономикалық көрсеткіштердің классификаторларын, өндіріс пен ауылшаруашылық өнімдер атауларын және т.б. атауға болады. Осындай терминологиялық сөздіктер ішінен әсіресе рубрикатор-сөздіктер маңызды рөл атқарады. Рубрикатор-сөздіктерді сала мамандары мен жеке мекемелер үшін ақпараттық қызмет ету жүйесіндегі әдістемелік құжат деп те атауға болады.

Рубрикаторлар, көбінде, өндірістік саланың негізгі бағыттарының жүйеленген ретін және маңызды нысандар мен процестерді қамтиды.

Салалық жиілік сөздіктердің өзіне тән маңыздылығы – кейбір бөлектелген терминдердің жиіліктерінің қамтылуы. Сөздіктің ең бір аса маңызды сипаты – терминдердің тақырыптық сипатта қамтылуы. Егер жиілік сөздік бірнеше пәндік аяны қамтитын болса, онда ол көпсалалық сөздік деп аталады. Көбіне, ғылымның дербес бөлініп шығуынан тарсалалық сөздіктер жасау қажеттігі туындайды. Мұндай сөздіктерді жасау тәжірибесінде түсіндірме терминологиялық сөздік типтерін құрастыру да аса маңызды. Түсіндірме терминологиялық сөздіктердің міндеті – нақты бір сала бойынша анықтама құралы болу және барлық терминдерді қамтумен бірге ұғымдар жүйесінің көрінісін беру. Мұндай мәліметтер мамандардың ғылыми және тәжірибелік жұмыстары үшін аса қажет деп саналады.

Терминологиялық сөздіктердегі мәліметтерді реттеуде дәстүрлі әліпбилік және ұялы әдістермен бірге статистикалық әдіс те қолданыс табады. Яғни мәліметтерді реттеуде терминнің мәтіндегі жиі немесе сирек кездесу сипаты да ескеріледі [10].

Әртүрлі тілдердің деректерін қамту жағынан терминологиялық сөздіктер біртілді және екітілді, үштілді және көптілді аударма сөздіктер деп бөлінеді [6; 7]. Түсіндірме сөздіктерді жасаудағы негізгі ұстаным – терминнің жаңа мағынаға ие болуы.

Терминологиялық сөздіктердің әртүрлі типтерінің негізгі дереккөзі ретінде жетекші ғалымдар мен маман-тәжірибешілердің қаламынан жарық көрген белгілі бір білім саласына қатысты жинақтап-қорытындылаушы монографиялар мен мақалалар материалдары алынады. Сонымен бірге қажетті материалдардың қатарына жоғары оқу орындарының ірі оқулықтарын да жатқызуға болады. Әдетте, тек осындай оқулықтарда ғана ғылыми ұғымдар мен терминдердің аса айқын және нақты анықтамалары беріледі. Энциклопедиялар мен басқа да сөздіктерде берілген салалық терминдердің барлығы бірдей ғылыми-техникалық терминологиялық сөздіктерде орын алуы қажет.

Ғылыми-техникалық терминологиялық сөздіктердің лексикалық қорына қосымша ретінде реферативті журналдардағы кітап мазмұнының рефераттары, жаңалықтары, патенттер мен іскағаздарының деректері де қамтылып алынады. Аса «сығымдалған» және шоғырландырылған түрде берілген мұндай мәліметтер негізгі терминдер жайлы жеткілікті түрде түсініктеме бере алады.

Терминологиялық сөздік түзу кезіндегі әртүрлі дереккөздерден таңдалып алынған терминдер негізінде терминологиялық сөзтізбе құрастырылады. Бұндай сөзтізбелердің реестрлік бірлігіне жататындар: сөз, сөз тіркес, қысқартулар, символдар және олардың тіркестері. Әдетте терминологиялық сөздіктің сөзтізбесі таза интуитивті тұрғыда түзіледі, яғни әрбір сала мамандарының жеке көзқарасындағы негізгі және маңызды деп аталатын терминдері таңдалып алынады. Әрине мұндай әдіс ғылыми әдебиеттерді талдауға негізделмеген, тек қана субъективті пікірлік ұстаным екені белгілі.

Ғылыми-техникалық терминологиялық сөздікті құрастырудан бұрын, маман-лексикограф ең алдымен баспаға дайындалып жатқан еңбектің хронологиялық шеңберін анықтап алуы қажет. Ғылымның жаңа салалары, мысалы, есептеу техникасы, радиофизика, биофизика үшін мұндай мәселе онша өзекті болмауы да мүмкін, себебі, ондай сөздікке енетін кез келген лексиканың уақыт мерзімі беріден басталады. Ал жаратылыстану ғылымдарының – физика, механика, математика, химия, биология, география, геология және көптеген қоғамдық ғылымдар салаларында мәселе басқаша. Бұл ғылымдардың орысша мәтіндері өзінің тарихи дамуында 150–200 жылдай мезгілді қамтиды. Осы себептерден терминологиялық сөздікте көрініс табатын терминдердің хронологиялық тәртібі маңызды мәселелер қатарына жатады.

Кейбір ғылыми кітаптар мен мақалаларда өткен кезеңдердегі ғалымдардың зерттеу нәтижелерімен пікірталас болып жатады. Мысалы, XIX-XX ғасырлардағы ғалымдардың зерттеулеріндегі ұғымдар мен тұжырымдамаларына көңіл бөліну жағы, көбінде, филология, математика,

биология, химия салаларына қатысты болады. Егер мұның бәрі тек қана табиғи тіл арқылы ғана іске асатынын ескерсек, терминологиялық сөздіктерге лексикалық қорды таңдап алу кезеңдік сипатта болатынына және тіл туралы ғылым тарихының өзі де ғылым тарихымен ұштасып жататынына көз жеткізуге болады.

Терминологиялық сөздіктердің сөзтізбесін қалыптастырғанда субъективті пікірді азайтудың бірден-бір жолы – нақты ғылым мен білім саласына тән логика-ұғымдық жүйені алдын ала модельдеу. Арнайы мәтіндерді логика-ұғымдық тұрғыда талдау тәсілі терминологиялық сөздіктерді жасаудағы алдын ала қойылатын қажетті шарт деуге болады.

Терминологиялық сөздіктердің құрылымы жайлы мынадай екі негізгі мәселені атауға болады:

1) терминологиялық сөздіктегі мәліметтерді қалыптастыру принциптері жайлы мәселе;

2) терминологиялық сөздіктегі жеке фрагменттердің құрылымы жайлы мәселе.

Бірінші мәселеге байланысты мынаны айтуға болады. Сөздік материалын ұйымдастыруға қатысты ғылыми-техникалық лексикографияда классикалық лексикографиядағы сияқты – әліпбилік, әліпбилік-ұялы, идеографиялық жалпы принциптер қарастырылады.

Терминологиялық сөздіктердің басым көпшілігі әліпби принципімен құрастырылатыны белгілі. Бұл ұстаным кез келген сөздікті пайдалану мен ондағы бірліктердің орналасу тәртібі жағынан көптен бері өзін-өзі ақтап келеді.

Осы айтылғандарға қарамай, мамандардың көптеген практикалық мұқтаждығынан аңғаратынымыз, ол терминологиялық сөздіктердегі терминдердің орналасуы тек әліпбилік немесе әліпбилік-ұялы принциптерге ғана сүйенбей, сонымен бірге логика-ұғымдық негізде, яғни ұғымдардың қатал ғылыми топтастырылуы бойынша идеографиялық принципте жасалуының қажеттігі. Бұл жердегі идеографикалық принципті қайсыбір ұғымды таңба арқылы белгілеп жазу деп түсіну керек. Мұндай принцип нақты ғылымның, білім саласының даму қисынына сәйкес келуімен бірге, терминді мағынасы бойынша сөздіктен тез тауып алуға да мүмкіндік жасайды. Мәселен, нақты ғылымдар мамандарына, көбінде, қайсыбір ұғымның қалай және қандай сөзбен аталатыны немесе ол сөздің синонимдерінің қандай ұғымдармен байланысы бар екендігін білу қажет болады. Маманның зерттеу тәжірибесінде және аударма практикасында белгілі бір терминді тұрпаты жағынан ұқсас болып келетін ертеректердегі зерттеулердің терминдерімен сәйкестендіру қажеттігі де туындауы мүмкін [11].

Идеографикалық принциппен құрастырылған терминологиялық сөздіктің негізгі құрамы – сол жағындағы терминдердің әліпби тәртібімен берілуі ғана емес, ол вербальды түрде берілген және шартты немесе формалданған тұрпаттағы ұғым болуы қажет. Аталған ұғым сөздіктің жалпы логика-ұғымдық құрылымында қатаң түрде өз орнына ие болады. Мысалы,

Роженің тезаурусында (*Roget's thesaurus…, 1978*) сөздік мақаланың басы ұғымның коды арқылы (үш орынды сан) және осы ұғымды білдіретін тірек сөзбен басталады. Тірек сөзден кейін сызықша және сөз табының белгісі қойылады, сосын сол сөздің синонимдері беріледі.

Ғылыми-техникалық терминологиялық сөздіктің жеке фрагменттерінің құрылымы. Лексикографияның жалпы теориясына сәйкес, сөздік мақала, көп жағдайда, мынадай бөліктерден тұрады [11]:

1) сөздік мақаланың аты;

2) терминдердің грамматикалық сипаттамасы;

3) терминнің қолданысына байланысты стильдік реңкі мен пайда болуына қатысты сипаттамасы;

4) терминнің семантикалық сипаттамасы.

Әліпбилік принципте түзілген терминологиялық сөздік мақаланың құрылымы түсіндірме және энциклопедиялық сөздіктердің құрылымына ұқсас болып келеді. Терминологиялық сөздіктерде сөздік мақаланың аты, әдетте, әліпби тәртібімен орналасады да, олар тек термин сөздер ғана емес, сөз тіркестер, сөздер мен символдардың комбинациясы, аббревиатуралық қысқартулар болуы да мүмкін.

Әліпбилік принциппен құрылған ғылыми-техникалық терминологиялық сөздіктерде терминнің семантикалық сипаттамасын беру оның мағынасының логикалық анықтамасы арқылы немесе аударма терминологиялық сөздіктердегі баламасы арқылы іске асады. Терминнің әрі шағын, әрі дәл логикалық анықтамасының берілуі, оның семантикалық құрылымын сипаттаудың ең бір тиімді жолы деп саналады.

Терминологиялық сөздіктерде терминдердің мағынасын анықтаудың ең бір дұрыс жолы, ол әрбір ұғымның типіне қарай, бірізділікті сақтай отыра, алдын ала белгілі параметрлерге сүйену. Терминнің толық түрдегі лексикографикалық анықтамасын құруға қатысты мынадай талаптарды ұстаным еткен жөн: логика-ұғымдық сызбаға алдын ала талдау жасау, ұғымның шығу тегін анықтау, ұғымға тән маңызды белгілерді анықтау, осы ұғымның басқа ұғымдар аясы ішінен ерекше белгілерін анықтау.

Арнайы әдебиетті аудару ісімен айналысатын аудармашыға қажетті сөздің баламасын табумен қатар оның синонимдерін де іздеп табу қажет болады. Яғни терминді мазмұндық тұрғыда кеңірек танып, сөздікте көрсетілген сөз, шындығында да, түпнұсқадағы ұғымды білдіру-білдірмеуіне көз жеткізу керек болады [6; 7].

Ғылыми-техникалық терминологиялық сөздіктерді құрастыруда терминнің анықтамасын тұжырымды етіп беру үшін мынадай ережелерді басшылыққа алған жөн саналады [11]:

1. Термин арқылы аталатын барлық ұғымдар мен нысандар анықтама ішінде қамтылуы қажет және, керісінше, анықтамада сөз болатын барлық ұғымдар мен нысандар терминнің атындағы ұғымдар мен нысандарға сәйкес келуі қажет.

2. Терминнің анықтамасында пайдаланылатын белгілер қарастырып отырған тақырыптық ая бойынша ең маңызды және мәнді болуы қажет.

3. Терминнің анықтамасында басқа бір терминнің анықтамасына сілтеме жасауға болмайды. Себебі, егер ол терминнің анықтамасында тағы да басқа бір терминнің анықтамасына жүгінсе, оның шегі табылмай, қайта оралып алғашқы терминге келуіміз де мүмкін.

4. Анықтамада пайдаланылған терминдер осы сөздік ішінде анықталған болуы немесе бұрыннан да белгілі және бірізді түсінігі болуы шарт. Егер анықтамада пайдаланылған терминдер іргелес жатқан басқа пәндерге қатысы болса немесе ол жалпыпәндік термин болса, онда арнайы қосымшаларға сілтеме жасалғаны жөн. Басқаша жағдайда ол терминнің түсінігін қандай әдебиеттен оқып-білуге болатыны айтылуы керек.

5. Терминнің нақты мағынасын ашатын анықтама берілуі үшін оның ғылыми ұғымдар жүйесіндегі алатын орнын жете түсіну аса маңызды.

6. Терминдердің анықтамалары, көптеген жағдайда, сөздермен қатар, математикалық және химиялық формулалар мен неше түрлі безендірілулер арқылы көрініс табады.

7. Әр елдің тілдеріндегі бір ғана терминнің мазмұнына қатысты анықтамалар бір-біріне сәйкес болмауы мүмкін. Сондықтан терминнің анықтамасын берерде басқа бір сөздікке сүйенетін болсақ, оның қай елдің сөздігінен алынғаны көрсетілуі керек.

Біртілді немесе аударма терминологиялық сөздіктерде терминнің семантикалық сипаттамасы синонимия, антонимия, текті-түрлік және ассоциативтік семантикалық қатынастар арқылы толықтырылуына әрі кеңейтілуіне болады.

Ғылыми-техникалық терминологиялық сөздікте нақты білім, ғылым, техника салаларының терминдік жүйесі көрініс табуы қажет. Ғылыми мәтіннің семантикалық мазмұнын дәл ашу, сол сала маманының қатысуынсыз мүмкін емес екендігі белгілі. Сондықтан ұғымның түсінігін сипаттау сол ғылыми сала маманының ісі де, ал оны барынша дәл және ықшам түрде тұжырымдап оқырманға жеткізу тілші-лексикографтың үлесі. Ал ғылыми-техникалық терминологиялық сөздіктің сапасы сала маманы мен тілші-лексикограф маманның бірлесе әрекет жасауының нәтижесі деуге болады.

Жоғарыда аталған арнаулы сөздіктерді құрастырудың теориясы мен практикасына байланысты тұжырымдар қазақ тілінің ғылыми-техникалық лексикографиясының дамуына оң әсерін тигізеді деген ойдамыз.

Терминологиялық деректер базасын құру бағыты – терминологиялық қызмет жасаудағы заман сұранысына сай, ең бір жаңа бағыт болып саналады. Мұндай бағытқа бастама болған терминологиялық кәртішкілерді жинақтау мен олардағы ғылыми, техникалық, экономикалық және т.б. пәндерге қатысты кәртішкілерді қажеттікке сай тауып алу жұмысы. Сонымен бірге, ондай кәртішкі көлемінің үнемі өсіп отыру сипатына қарай қажетті ақпаратты тауып алу мен оларды өңдеу жұмыстарының барынша қиындауы да негізгі себептердің бірі деуге болды. Аталған жұмыстарды жеңілдету үшін ондай іс-әрекеттерді компьютерге жүктеу қажеттігі туындады. Себебі қазіргі кезде аса көлемді деректерді сақтау мен оларды

өңдеу және қажетті деген ақпаратты тез арада іздеп-табу жұмыстарын тек компьютерге жүктеу арқылы іске асыруға болатыны белгілі. Бұл бағыт бойынша атқарылатын жұмыстың маңыздылығы – мамандар мен ғалымдарды нақты және нормативті терминологиялық құжаттар арқылы бекітілген деректермен дер уақытында қамтамасыз ету.

Тілдің терминологиялық деректер базасын (ТДБ) жасау ісінің теориялық та, практикалық та болашағы зор деп саналады. Ол автоматтанған «қойма» болудан әрі қарай дамып, аса маңызды ғылыми-зерттеу құралына айналып отыр.

Ұйымдастыру және әдістемелік шаралар. Тақырыпта аталған шаралар түрлері мыналар: терминологиялық деректер базасын (ТДБ) құруда терминологиялық стандарттар мен сөздіктерді түзуге жетекшілік ету, терминология мәселесі жайлы ғылыми конференцияларды, семинарларды және т.б. ғылыми кездесулерді өткізу, әртүрлі мамандар категорияларын (қолданбалы лингвистика аясындағы аудармашылар, стандарттаушылар, ғылыми қызметкерлер) терминтану негіздеріне оқыту. Осы аталғандар және басқа да ұйымдастыру-әдістемелік қызмет түрлерін ұлттық және халықаралық ұйымдар атқарады.

Тақырыпты түйіндей келе, мынадай қорытынды жасауға болады: қазіргі кезде бүкіл әлем бойынша теориялық және практикалық тұрғыда терминологиялық қызметке барынша көңіл бөлінуде. Өйткені кез келген елдің екінші бір мемлекетпен жасалатын қарым-қатынасының басты құралы тіл болып табылады және тілдік деңгейлердің ішінде терминдік бірліктер негізгі рөл атқарады. Осы ретте Тәуелсіз Қазақстан Республикасының мемлекеттік тілі – қазақ тілінің терминологиялық қорын қалыптастыру, жүйелеу, оны автоматтандыру және т.б. қолданбалы қазақ тіл білімінің өзекті мәселелерінің бірі болып отыр. Мұның себебі – ғалымдар мен өндіріс мамандарының рационалды және дәл сипаттағы терминологиясыз ғылым мен техникада, коммерция мен экономикада айтарлықтай жетістікке жету мүмкін емес.

Деректер базасындағы ақпараттық ағын. Әлем мемлекеттерінің, оның ішінде Тәуелсіз Қазақстан Республикасының ғылымы мен техникасының даму деңгейі көбінде ғылыми-техникалық ақпараттардың қаншалықты дер кезінде және сапалы түрде өңдеу дәрежесімен анықталады. Адам баласының әрдайым дамып және қарқындап өсу сипатындағы ақпарат ағымын меңгеріп алу мүмкіндігі тек жаңа технологияны қолдану арқылы ғана жүзеге асады. Осыған қатысты мәтінді өңдеу тұсында қажетті ақпаратты іздестіру мәселесі туындайды. Ал ол үшін қажетті ақпарат тілдің деректер базасында жинақталуы тиіс.

Деректер базасы ұғымын аса сыйымды қоймамен барабар деп ұғынуға болады. Мысалы, бір мекеменің (фирманың, зауыттың) өндіретін өнім деректерінің бәрі бірдей сол қоймада сақталады және әртүрлі тұтынушы өзіне қажетті деген мәліметті сол қоймадан алып пайдалануына мүмкіндік туындайды. Барлық деректерді есте сақтайтын жад құрылғысы бір жерде немесе бірнеше жерде орналасуы да мүмкін. Соңғы жағдайда деректерді бір

жерден екінші жерге жеткізіп тұратын қосымша құралдармен қамтамасыз етіп тұру қажет болады. Егер деректер мәлімет түрінде болатын болса, онда оған бағдарлама арқылы еркін қатынас жасауға (кіруге) мүмкіндік тууы керек. Яғни бұл жағдайда қойма деп отырғанымыз компьютер жадында арнайы ұйымдастырылған мәліметтер жиыны деуге болады. Ал бір ғана жерге біріктірілген тұтастық сипаттағы деректер базасы өте күрделі және тіпті шындыққа жанаса бермейтіндей де көрінеді. Осындай қиындықтарға қарамай, аталғандай күрделі деректер базасын құрастыру қажет-ақ және оларды кезең-кезеңге бөліп жүзеге асыру керек болады. Шындығында, бүгінгі күнгі жүзеге асырылған көптеген деректер базасы шектелген тәжірибелік қолданысты көздейтін мәселелерге арналған. Көпшілік жағдайда бір компьютерде бірнеше деректер базасы орналасады. Уақыт өте олардың ішіндегі бір-біріне қызметі ұқсас болып келетін базаларды біріктіруге де болады. Әрине, егер ондай әрекеттен тек база жұмыстарының тиімділігі мен қарқындылығы артатын болса ғана базаларды біріктіру жүзеге асады.

Деректер базасына өзара байланыста болатын және бір қоймада сақталатын деректер жиыны деп те анықтама беруге болады [12, 28-б.]. Сонымен бірге олардың араларында бір-біріне ең аз кедергі келтіретін басы артық ақпараттың болғанымен, бір немесе бірнеше мәселені бірге шешуге қолдануда ондай жүйе қолайлылық пен нәтижелікке ие болады; деректердің компьютер жадында орналасуы, оларды пайдаланатын бағдарлама түрінде тәуелсіз болады; базада бұрыннан бар деректерге жаңадан мәлімет қосу, яғни толықтыру немесе оларды өзгертуге және қажетті деректі іздеп табу үшін бәріне ортақ басқару тәсілі қолднылады. Деректердің компьютер жадында орналасу құрылымы болашақта шешілетін қолданбалы мәселелер санын ұлғайтуға мүмкіндік тудырады. Егер бірнеше деректер базасы құрылымдық жағынан толығымен дербестік шартты қанағаттандырса, онда мұндай жүйе деректер базасының жиынтығынан тұрады деп те айтуға болады.

Қарапайым ұйымдасқан деректермен әрекет ететін жүйелер үшін әрбір шешілетін мәселенің өзіне тән «жазба» жиынтығы болады. Деректер базасының міндеті – бір ғана деректер жиынтығын сан жағынан барынша көп болып келетін қолданбалы мәселелерді шешу. Сондықтан деректер базасын көптеген пайдаланушылардың қажеттігін өтейтін ақпарат қоймасы ретінде де құрастырады. Мәселен, оларға осы деректердің қажеттігі бар зауыт, басқару мекемелер немесе басқа да қызмет орындарын жатқызуға болады. Мұндай деректер базасы сол мекеменің басқару әрекетіне қажет болатын ақпаратпен ғана қамтамасыз етіп қоймай, деректерге әрдайым өзгеріс енгізіп (модификациялап) тұруы қажет. Сонымен бірге, мысалы, жобалау жұмыстары үшін немесе тиісті сауалдарға тиісті жауап беруге қажетті ақпаратты деректер базасынан іздеп табу қажет болады. Деректер жиынтығымен, араларында ведомстволық кедергілер болуына қарамай, олармен бірнеше мекемелер де пайдалануына болады.

Деректер базасы деректерді дестелік (пакеттік) өңдеу қажеттігі үшін, реалды уақыт ішінде немесе жедел мезгілде (қолма-қол) өңдеу қажеттігі үшін

де құрастырылуы мүмкін. Аталған өңдеу әдістері көптеген деректер базаларында ескеріледі, ал басқа да деректер базасы жүйелерінде ақиқат уақыт ішінде терминалдарға қызмет ету деректерді дестелік өңдеу жұмысымен бір мезгілде жүзеге асыруға мүмкіндік бар.

Ақпарат артықтығы жайында. Деректер базасы дискілік немесе таспалық «кітапханалардың» басым көпшілігі қайталанатын ақпаратты көптеп сақтайды. Деректердің көпшілік элементтерін сыртқы есте сақтау құралдарына жазып алу әрекетінде «артық ақпарат» сақтауға рұхсат етіледі. Себебі оларда бірдей дәрежедегі деректер әртүрлі мақсатқа сай және модификацияланған бірнеше варианттары жазылып алынады. Ал деректер базасы, айтарлықтай дәрежеде, мұндай ақпарат артықтығынан «құтқаруға» мүмкіндік жасайды. Сондықтан деректер базасын кейде деректер элементтерінің «артықсыз жиынтығы» деп те атайды. Бірақ та деректерге «қатынас жасау» (доступ) мезгілін азайту үшін немесе адрестеу тәсілдерін жеңілдету үшін көптеген деректер базасында (аздаған дәрежеде болса да) артық деректер болатындығы шындық. Кейбір жазбалардың қайталануы, олардың кездейсоқ жоғалып кету жағдайында алғашқы қалпына келтіру үшін де қажет болады. Деректер базасының (ДБ-ның) ақпараты артықсыз болуы үшін кейбір жағдайларда ымыраға келуге тура келеді. Мұндай кездерде жақсы басқарылатын немесе ең аз көлемдегі ақпарат артықтығы жайлы, немесе жақсы құрастырылған (зерттелген) деректер базаларда қажетсіз артық ақпарат болмайды деп санайды [12, 29-б.].

Басқаруға келмейтін артықтың бірнеше кемшілігін атауға болады. Біріншіден, деректердің бірнеше көшірмесін сақтау қосымша шығынды қажет ететіні белгілі. Екіншіден, (аса мәнді) бірнеше артық көшірмелерді жаңартып отыру үшін де бір амалдың өзін қайта-қайта қайталауға тура келеді. Сондықтан егер файлдарды өңдеу кезінде көп көлемді ақпарат жаңартылатын болса немесе жаңа элементтерді енгізу мен ескіні жою әрекеті де жиі қайталанатын болса, ақпарат артықтығы барынша «қымбатқа» түседі. Үшіншіден, деректердің әртүрлі көшірмелері жаңартудың басқа кезеңдердегі ақпаратпен сәйкес келіп жатса, олар кейбір қайшылыққа әкелуі мүмкін.

ДБ-ны қолданбаған жағдайда, аса көлемді ақпаратты өңдеу кезінде көптеген артық деректер пайда болады да оларды базаның жадында сақтау мүмкін болмай қалады. Деректердің айқын түрдегі қайшылықтары компьютерден алынған ақпаратқа сенімсіздік тудырады.

Көптеген деректер базаларының ең маңызды сипаттамаларының бірі – *үздіксіз өзгеріс* пен *ұлғайтуды* қажет етуі. Деректердің жаңа түрлері қосылған сайын деректер базасының құрылымы аса жылдамдықпен өзгертіліп тұруы қажет. Жүйені талдаумен айналысатын мамандар үшін таңдалып алынған деректер құрылымына өзгеріс енгізу қажеттігі туындаса, онда қолданбалы компьютерлік бағдарламаларды қайта жазуға немесе жөндеуге тура келеді. Бірақ ондай өзгерістердің барынша аз болу жағына көбірек мән беру керек болады. Қолданбалы бағдарламалар саны неғұрлым көп болса, соншалықты деректер құрылымына өзгеріс енгізу шарасы да

күрделінеді. Деректер базасы өзгерісінің қарапайымдылығы, оның өндірісті басқарудағы қолданбалы мәселелердің даму процесіне әсерін тигізері сөзсіз.

Деректер базасының негізгі қасиеттерінің бірі – деректердің тәуелсіздігі. Мұндай тәуелсіздікті сөз еткенде, оның астарында қолданатын қолданбалы бағдарламалардың да бір-бірімен тәуелсіздігі жатқанын ескеру қажет болады. Ал бұл тәуелсіздікті – бір бағдарламада орын алатын өзгеріс, басқа бағдарламаларға өзгеріс әкелмейді деп ұғыну қажет болады. Шындығында, деректердің толығымен тәуелсіздігі және сол сияқты олардың артық ақпаратсыз болуы да өте сирек кездесетін жағдай [12, 31-б.].

Егер деректер элементтерінің бір ғана жиынтығы бірнеше қолданбалы мәселелерді шешуге қатысатын болса, бұл жиынтық элементтерінің арасында қолданбалы бағдарламаларға қажетті көптеген әртүрлі өзара байланыс орнатылуы қажет болады. Деректер байланысын ұйымдастыру деректер мен жазбалар элементтері арасындағы байланысты жүзеге асыруға да қатысты болып келеді. Көптеген қолданбалы мәселелерді шешуге қатысты қолданыс табатын деректер базасында деректер арасындағы көптеген аралық байланыстар орнатылады. Аталған жағдайда деректерді жүйе жадында ұзақ мезгіл сақтау мен қолдану әрекетінде, олардың дұрыстығын қадағалау және қорғау мен құпиялығын қамтамасыз ету әрекеті, өзара тәуелсіз қарапайым файлдардағы осындай әрекеттерге қарағанда қиынға түседі. Сондықтан деректердің құпиялығын сақтау мен компьютердің «жаңылысу» (сбой) мүмкіндігін қайтадан қалпына келтіру мәселесі ДБ-ны құрастыруда аса маңызды болып саналады.

«Деректер базасы» ұғымы өткен ғасырдың 60-шы жылдарында пайда болғаны белгілі. Оған дейін деректерді өңдеу саласында деректерге байланысты «файл» мен «жиынтық» жайлы ғана сөз болатын. Жаңа ұғым (деректер базасы) үлгіге айналғаннан кейін, көптеген пайдаланушылар «база» сөзін «файл» сөзінің орнына қолданып, ал бірақ базаға ғана қатысты мынадай қасиеттерді: *артықсыз, тәуелсіз, деректердің өзара байланыстылығы, деректерді қорғау* немесе *ақиқат уақыт аралығында деректерге қатынас жасау мүмкіндігі* сол күйінше қалдыруды да әдетке айналдырғаны мәлім. Деректер базасын бағдарламалық қамтамасыз етудің барынша дамыған түрінің қолданысқа енуінен бастап, аталған қасиеттер қайтадан өз мағынасына ие болды.

Деректер базасы қызметінің тиімділігін көтеру мақсатында және жаңа «сұратулардың» қосылып тұруынан ДБ-ның құрылымы да өзгеріске ұшырды. Пайдаланушылар ДБ жүйесінде қойылатын талаптарға өзгерту енгізумен бірге деректерге қойылатын сұрату типтерін де түрлендірді.

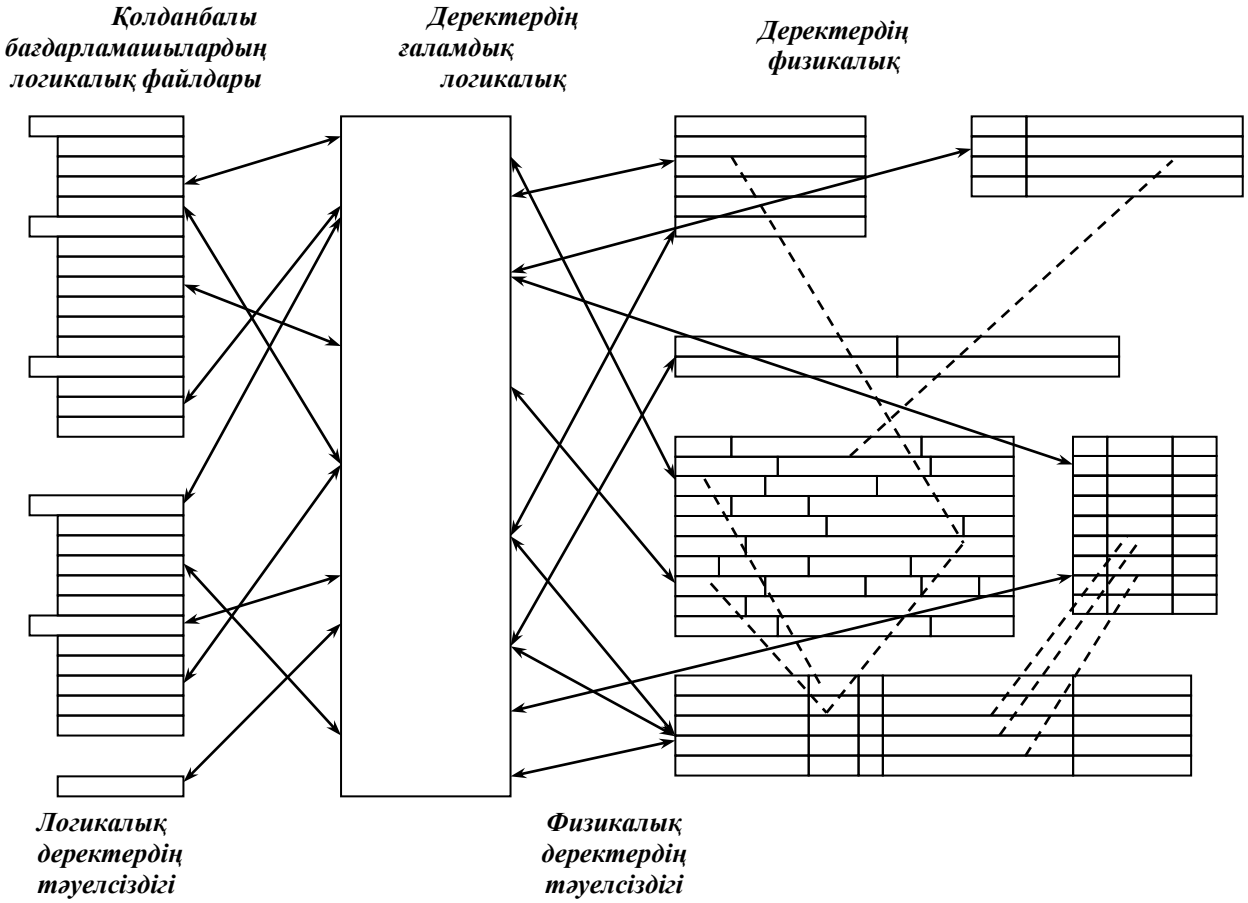
Деректер базасы құрылымының статикалық сипаты файлдық құрылымға қарағанда төменгі дәрежеде болуымен бірге, жүйе жадында сақталатын деректер элементтері мен оларды жадта сақтау тәсілдері де үнемі өзгеріске ұшырап отырады.

Бір ғана деректерден әртүрлі қисынды (логикалық) файлдар алынуы мүмкін, ал әртүрлі қолданбалы мәселелерге қатысты бір ғана дерекке «қатынас жасау» сол қолданбалы мәселелердің талаптарына жауап беретін

әртүрлі жолдармен жүзеге асады [12, 37-б.]. Бұл жағдай деректердің күрделі құрылымын құрастыруға әкелетіні белгілі. Бірақ деректер базасын жоғары дәрежеде бағдарламалық қамтамасыз ету жұмысы қолданбалы бағдарламашыны күрделі құрылымға байланысты қиындықтардан құтқарып алуға барынша жол ашты.

Бағдарламалық қамтамасыз етудің деңгейіне қарай, қолданбалы бағдарламашы белгілі бір жағдайды өзіне қажетті деректер элементінің немесе «жазбаның» тек атын ғана хабарлауы жеткілікті болады. Ал басқа бір жағдайда (мысалы, бағдарламалық қамтамасыз ету жағдайында) деректер элементінің теңдестіру (идентификациясын) мен жиынтық атын да хабарлауы тиіс.

Деректердің логикалық және физикалық тәуелсіздігі. Деректердің жалпы логикалық құрылымы, әдетте, аса күрделі және деректер базасының ұлғаю сипатына қарай, үнемі өзгеріске ұшырап отырады. Сондықтан жүйеде қолданылатын көптеген қолданбалы бағдарламаларды өзгеріссіз қалдырып, жалпы логикалық құрылымды ғана өзгертуді қамтамасыз ету аса маңызды. Кейбір жүйелерде деректердің жалпы логикалық құрылымын өзгерту, оның әрекет ету формасын білдіреді, яғни бұл құрылымның үнемі даму сипатында болатындығының айғағы.



4.1-сызба

Сондықтан деректердің тәуелсіздігінің екі деңгейінің болуы талап етіледі. Олар деректердің *логикалық* және *физикалық тәуелсіздігі* деп аталады [12, 38-б.].

Деректердің логикалық тәуелсіздігі жағдайында жалпы логикалық құрылым өзгеріске ұшырауы мүмкін, бірақ қолданбалы бағдарламалар өзгеріссіз қалулары керек. Ал деректердің физикалық тәуелсіздігі жағдайында олардың физикалық орналасуы мен ұйымдасу жақтары өзгеріске ұшырауы мүмкін, бірақ ол деректердің жалпы логикалық құрылымы мен қолданбалы бағдарламаларына ешбір өзгеріс әкелмейді.

4.1-сызбада деректердің логикалық және физикалық тәуелсіздігі жайлы ой-пікір (идея) безендірілген. Аталған суреттегі ортада орналасқан үлкен тік төртбұрыш деректердің жалпы логикалық құрылымын көрсетеді, ал кейбір жағдайда, ол деректердің ғаламдық логикалық көрсетімі деп те аталады. Бұл көрсетім деректердің физикалық құрылымынан және олардың нақты қолданбалы бағдарлама көрсетімінен, айтарлықтай дәрежеде, айырым табуы мүмкін. Деректер базасының бағдарламалық қамтамасыз етілуі жалпы логикалық құрылымнан қолданбалы бағдарламашының деректер көрсетімін өзіне қабылдайды, ал сосын барып, оны жалпы логикалық құрылымды деректердің физикалық көрсетіміне бейнелейді.

Қазіргі кездегі деректер базасына қойылатын талаптар ***4.1-сызбадағыдай*** көрініс тапты және олар мыналар [12, 35-б.]:

- бағдарламалық құралдар деректердің логикалық және физикалық тәуелсіздігін қамтамасыз етуі;
- деректер базасының даму барысының көп шығынды қажет етпеуі;
- деректер әкімшілігі үшін қарастырылған құралдардың басқару жүйесінің қызметін орындауы және деректердің сақтығын қамтамасыз етуі;
- деректерді қорғау, құпиялығын сақтау, тұтастығы мен қауіпсіздігін басқару процедураларының тиімділігін қамтамасыз етуі;
- кейбір жүйелерде деректерде базадан тез тауып алу үшін инвертирленген (дағдылы реттігі өзгерген) файлдарды қолдану мүмкіндігінің болуы;
- алдын ала жоспарланған ақпараттық сұратуларға жауап беруді де көздейтін деректер базасын құрастыру;
- деректер орнын жылжыту мүмкіндікке ие құралдармен қамтамасыз ету;
- деректер базасының әкімшілігі үшін деректерді сипаттау тілінің, қолданбалы бағдарламашы үшін бұйрықтар тілінің (язык команд) және, кейбір жағдайларда, пайдаланушылар үшін сұраулар тілінің (язык опросов) болуын бағдарламалық қамтамасыз етуде есекру.

Деректердің концептуалдық моделі. Қазіргі кездегі деректер базаларының құрылымындағы *ортада орналасқан тіктөртбұрыш* аса маңызды элемент деп саналады (***4.1-сызбаны*** қара). Ол жүйедегі деректердің концептуалдық көрсетімін белгілейді. Кейде оны *деректер моделі* немесе *концептуалды модель* деп те анықтайды. Ал сонымен бірге, оны *сызба* немесе *концептуалдық сызба* деп те атауға болады. Концептуалдық сызбаны,

мүмкіндігінше, тұрақты етіп құрастыру қажет болады. Деректерге қатысты бағдарламашылардың түсінігі өзгеріске ұшырағанда деректерді *физикалық сақтау* мен оларды *ұйымдастыру* жағдаяты да өзгереді. Бірақ бұл жағдайда *концептуалдық модель* өзгеріссіз қала береді немесе ол тек деректердің қосымша типтерін өзіне қосып алу мақсатында ғана ұлғаюы мүмкін. Деректердің концептуалдық моделі болашақта деректерді өңдеу ісінің берік іргетасындай мықты болуы керек. Концептуалдық модельдің тұрақтылығын қамтамасыз ету үшін оны деректердің ішкі қасиеттерін ескере зерттеп алып, құрастыру қажет болады. Енді осы жайтқа қысқаша тоқталайық.

Егер концептуалдық модель төменгі дәрежеде құрастырылса, онда оны пайдалану қажеттігімен толықтыру үшін модельді барынша жетілдіру керек болады. Ал бұл жағдай осы модельге деп арналып жазылған бағдарламаларды қайтадан жазуға мәжбүр етеді. Деректер базасын жобалаудағы, негізгі мәселе – жүйені дамыту барысында бағдарламаны қайта жазуға қатысты іс-әрекетті болғызбау.

4.1-сызбада көрініс тапқан құрылымның міндеті деректер базасын қалыптастыру мен оны пайдалануға қатысты жұмыстарға өзгеріс жасамай, деректер құрылымына өзгертулер енгізуге барынша еркіндік беру.

Аталған сызбада деректерді ұйымдастырудың үш типі көрініс тапты. Олар *сыртқы ұйымдасу*, *ғаламдық логикалық ұйымдасу* немесе *концептуалдық модель* және *физикалық ұйымдасу* деп аталып жүр. Бұл үш түрлі ұйымдасу типтері, әдетте, бір-бірінен қатты айырым табады [12, 40-б.].

1-тип. Деректердің сыртқы ұйымдасуы – қоданбалы бағдарламалардың немесе ақырғы пайдаланушының түсінігіндегі көрсетім.

2-тип. Деректердің ғаламдық логикалық ұйымдасуы – деректер базасының жалпы түрдегі ұйымдасуы немесе оның концептуалды моделі. Осындай ұйымдасудың негізінде әртүрлі ұйымдасу түрлері қалыптасуы мүмкін. Деректердің мұндай логикалық көрсетімі *деректердің физикалық ұйымдасуынан* толық сипатта тәуелсіз деп саналады. Мұндай логикалық көрсетім деректер базасын басқарудың бағдарламалық құралдарының бір бөлігі деп саналатын деректерді сипаттау тілінде баяндалады.

3-тип. Деректердің физикалық ұйымдасуы – деректердің физикалық көрсетімі және оларды есте сақтау құралдарында орналастыру. Бұл жазбаларды іздеуде қолданылатын физикалық құралдарға тәуелді (индикаторлар, көрсеткіштер, тізбелер және т.б.) және сонымен бірге «асыра толтыру аймағының» (область переполнения) орын алуымен және жаңа жазбаларды қосу мен ескілерді жою құралдарының болуымен анықталады.

Деректер базасын құрылымдық тұрғыда қарастырғанда әдетте оны үш түрге бөледі: **иерархиялық, желілік** және **реляциондық**.

Деректік базаның құрылымы «шежірелік» (бұтақ таралымы) принципіне негізделсе, оны **иерархиялық құрылымдық деректер базасы** деп атайды. Егер деректік базаның құрылымы еркін желі құрылымындай болса, онда оны **желілік құрылымдық деректер базасы** деп санайды. *Иерархиялық құрылымдағы деректер базасына* өңдеу жұмыстарын жүзеге асыру үшін жан-жақты зерттелген математикалық құралдар кең түрде

қолданылады. Ал желілік құрылымдағы деректер базаларымен әрекет ету соңғы кезде көптеген қиындықтарға әкеліп жүр. Сондықтан желілік деректер базасына қарағанда иерархиялық деректер базасы бойынша іздестіру жұмысын атқару аса нәтижелі болатындығы дәлелденді. Бірақ кез келген ақпаратты *иерархиялық құрылымда* бейнелеу ойдағыдай бола бермейтінін де ескеру қажет. Сол себепті егер ақпаратты баяндау тәсілі иерархиялық сипатта болса ғана іздестіру, өңдеу жұмыстары иерархиялық деректер базасына қолданылуы керек. Соңғы кезде жиі қолданыс тауып жүрген реляциялық деректер базалар тікбұрыштық кестелердің жиыны түрінде көрініс табады. Ал кестелер арасындағы байланыс – айқыш сілтемелер арқылы жүзеге асады.

Деректер базасы бойынша жұмыс істеу мүмкіндігіне ие болу үшін арнайы жазылған бағдарламалар қажет. Деректер базасындағы жұмысты қамтамасыз ететін бағдарламалар топтамасын **деректер базасын басқару жүйесі** деп атайды. Деректер базасын басқару жүйесінің құрамы әртүрлі деректер базалармен жұмыс істеуге мүмкіндік тудыратын барынша әмбебаптық сипаттағы бағдарламалардан тұруы қажет деп саналады.

Деректер базасын басқару жүйесі деректер базасының құрамдық жағын құрастыруды, деректер базасындағы ақпараттық өзгерісті (модификацияны), деректер базасындағы ақпаратты іздестіруді, тұтынушыға ақпараттың қолайлы түрде шығарылуы мен басқа да қосымша қызметтерді қамтамасыз етуді көздейді.

Деректер базасын құрамдау. Деректер базасын зерттеп, іске қосу барысында оның құрамдық жағына көп көңіл бөлінеді. Ақпаратты сақтау мен өңдеу нәтижелі болу үшін қандай ақпарат түрін сақтаудың қажеттігі, деректер базасында қандай нысандардың типтері орын алатындығы алдын ала айқындалуы қажет. Әрбір нысан типтері үшін оларды айқындайтын атрибуттар топтамасы анықталуы керек болады. Сонымен бірге әртүрлі типтегі нысандар арасында мүмкін болатын байланыс типтері де анықтауды қажет етеді. Барлық нысандардың типтері бейнеленген және әрбір нысан типінің атрибуттар топтамасы мен мүмкін болатын байланыс типтері көрініс тапқан сызбаны *жалпы концептуалдық сызба* немесе *деректер моделі* деп атайды.

Әдетте, деректер базасын зерттеу онда бейнеленетін ақпаратқа қатысты пәндік аяны зерттеуден және жалпы концептуалды сызбаны бейнелеуден басталады. Деректер базасын басқару жүйесінде жалпы концептуалдық сызбаның негізінде құрастырылатын деректер базасының құрылымы жайлы ақпаратты енгізу мүмкіндігі ойластырылуы керек. Осындай жағдайда әрбір атрибуттың деректер типін анықтау қажеттігі (бағдарламалау мағынасында) туындайды.

Әдетте деректер базасын құрастырғанда артықтық сипаттағы ақпараттың оның жадында сақталмағаны жөн деп саналады. Сондықтан, кейбір жағдайда, деректер базасын артықсыз (неизбыточная) деректер жиыны деп анықтайды. Бірақ іс жүзінде мұндай нәтижеге жету оңай емес. Көбінде ақпаратты іздестіру жұмысының ыңғайлы болуы үшін ақпараттың артықтығы қолайлы деп те саналады. Мысалы, *ұшақ рейсі* деп аталатын

нысанның мынадай атрибуттары *ұшу уақыты, қону уақыты* және *әуеде болу уақыты* ішінен екеуі ғана белгілі болса, үшінші атрибутты әр уақытта есептеп шығаруға болады. Бірақ егер біз деректер базасынан ақпаратты іздестіру әрекетін *ұшақ рейсі* сипаттамаларының әрбіреуі бойынша жүргізу қажет деп санасақ, үш түрлі атрибуттың да орын алғаны жөн болады. Солай болғанның өзінде де, жалпылама жағдайда, ақпарат артықтығының аз мөлшерде болуына талпынуымыз қажет.

Деректер базасындағы ақпараттың артықтық сипатының қауіптілігі тек қосымша жад көлемін қажет етуде ғана емес, ең алдымен, деректер модификациясының өзгеріске ұшырау салдарынан қадағалауға қажетті қосымша бақылау әрекетінің болуында. Сондықтан зерттеушілердің тұжырымдауынша, деректер базасындағы артықтықтың тек бақылау жұмыстарына ғана қатысты болғаны жөн деп саналады.

Деректер базасындағы ақпаратқа өзгерістер енгізу (модификациялар). Деректер базасын басқару жүйесінің базада орналасқан ақпаратқа тиісті өзгерістер енгізу мүмкіндігі болуы қажет. Ондай өзгерістердің типі мынадай: *жаңа жазбаны* (новая запись) немесе *жаңа нысанды қосу, бар жазбаны өшіру, бар жазбаның атрибут мәнін өзгерту.* Деректер базасының құрылымына модификация жүргізу ең күрделі өзгеріс түріне жатады. Іс жүзінде деректердің жалпы концептуалдық сызбасының идеалды түрін (өзгеріске жатпайтын) алдын ала болжамдау мүмкін бола бермейді. Сондықтан деректер базасын қолдану кезінде кейбір нысандар қажетсіз, ал қайсыбірінің жетіспеуі немесе артық болуы, тіпті кейбір тәуелділіктерді басқаша түрде ұйымдастыру қажет болады. Деректер базасын басқару жүйесінде деректер базасының құрылымын өзгерту тұсында ондағы қажетті деген ақпаратты сақтап қалу мүмкіндігі де ескерілуі қажет.

Ақпаратты іздестіру. Деректер базасын басқару жүйесі қызметінің ішінде ең бір маңызды деп саналатыны – пайдаланушының сұранысы бойынша жүргізілген іздестіру әрекеттерінің нәтижелі болуын қамтамасыз ету. Іздестіру жұмысы дәл және лезде орындалуы қажет деп ұйғарым жасалады. Ал ол үшін деректер жиынының белгілі пішімде ұйымдастырылуы қажет болады. Реттелген тізбе деректері бойынша іздестіру әрекетінің ойдағыдай нәтижелі болатындығы белгілі. Ал егер деректер тізбесі ретсіз түрде орналасса, онда қажетті деген нысанды тауып алу үшін барлық деректер жиынын толығымен, яғни басынан аяғына дейін қарастырып шығуға тура келеді. Мұндай әрекетке көп уақыт кететіні белгілі, ал уақытты үнемдеу үшін іздестіруді әртүрлі атрибуттардың мәні негізінде жүргізген жөн болады. Бұл мүмкіндік әдетте тек инверсияланған (дағдылы реті өзгерген) индекстік файлдарды құрастыру арқылы іске асады. Ондай файлдарда атрибуттардың мәні мен қажетті деген жазба түрлеріне сілтеме жасайтын арнайы түрдегі көрсеткіштер орын алады. Артықтық сипаттағы ақпаратты индекстік файлдар өз бойында сақтайды. Сондықтан мұндай ақпараттық артықтықты бақылау үшін деректер базасын басқару жүйесінде ақпаратты өзгертуге (модификациялауға) қажетті арнайы механизмнің орын алғаны жөн. Әрбір атрибутты олардың мәндері бойынша тез тауып алу үшін

оның өзіне ғана сәйкес келетін, индекстік файлы болуы қажет. Яғни деректер базасын басқару жүйесінде ондай міндетті атқаратын арнайы бағдарлама құрастырылуы керек деген сөз. Мәні бойынша атрибутты тез іздестіру мүмкіндігін оның «кілті» немесе *кілттік атрибуты* деп атайды. Жазбалар файлының реттелуіне қатысты «кілтті» *бірінші реттегі кілт,* ал басқаларын *екінші реттегі кілттер* деп ажыратады. Әдетте *бірінші реттегі кілттің* өзіндік артықшылыққа ие болуына немесе жазбаның бірмағыналы теңестірулерге мүмкіндігі болуына талпыну керек болады. *Екінші реттегі кілттердің* мәндері әртүрлі жазбаларда қайталанып тұруы мүмкін. Сондықтан *бірінші реттегі кілт* бойынша іздестіруді ұйымдастырудың жылдамдығы аса жоғары, яғни ол максимум шамаға ие. Деректер базасын басқару жүйесі іздестіру жұмысын кілттік атрибуттардың мәні бойынша ғана емес, сонымен бірге кілтсіз атрибуттар негізінде де іске асыра алады. Бірақ соңғы жағдайда іздестіру жылдамдығы кемиді. Деректер базасында ақпаратты іздестіруге қатысты қарапайым екі типтегі сұратулар орын алады. Бірінші тип – айқын атрибуттың мәнін (шамасын) белгілі нысан бойынша (бірінші реттегі кілт) іздестіру. Егер сұрату түрі күрделі болатын болса, ондайларды бірнеше қарапайым сұратуларға бөліп қарастыруға болады.

Деректер базасын басқару жүйесінің қосымша функциялары. Әдетте деректер базасын басқару жүйесі бірнеше қосымша қызметтерді қамтиды. Мысалы, ол көбінде іздестіру негізінде табылған ақпаратты пайдаланушыға қолайлы түрде (мысалы, кесте түрінде) – шығарып беруді ұйымдастыра алады. Сонымен бірге деректер базасын басқару жүйесі көптеген есептеу жұмыстарын да орындай алады. Мәселен, мынадай сұратуды: «деректер базасында белгілі шарттарды алдын ала қанағаттандыратын қанша нысан бар?».

Деректер базасын басқару жүйесінде басқа да есептеу жүйелерінің атқаратын кейбір функцияларын өз міндетіне алатын мүмкіндігі болуы қажет. Мысалы, оған жататындар: ақпаратты жүйелік «жаңылысу» (сбой) кезінде пайда болатын қателерден, тұтынушының кездейсоқ қателесуінен және қажеттікке қарай кейбір жүйеге рұхсатсыз кіруден сақтандыру жұмыстары.

Деректер базасының теориясы автоматтанған фактографиялық ақпарат іздестіру жүйесін құрудың математикалық және программалық құралдары ретінде көрініс табатынын да айтуға болады.

Деректер базасының әкімшілігі. Деректер базасының құрылымын жобалау және онымен жұмыс жүргізу үшін **деректер базасының әкімшілігі** деп аталатын жаңа ұғым енгізілді [12, 40-41-бб.].

Деректер базасының әкімшілігі мекеменің барлық деректерін немесе тек өз жүйесіне қатысы бар бөлігін ғана қорғайды. Сонымен бірге ол деректер базасының барлық құрылымын бақылап отыруды да жүзеге асырады.

Әкімшілік деректердің ұйымдасуына байланысты ғаламдық түсінікке ие болуынан, оған өте жиі назар аударуға тура келеді. Шындығында, деректер базасының әкімшілігі дегеніміз, көпшілік жағдайда, бір адам емес,

бір бөлім қызметкерлері немесе бір топ адам деуге де болады. Себебі деректер базасының табиғатын, оның ұйымдасуын, экономикасын, зерттеу критерийлерін және көптеген пайдаланушылардың талаптарын терең түсіну үшін бір ғана адам мүмкіндігінің аздық ететіні белгілі.

Қолданыстағы терминологиялық деректер қорының көпшілігі ғылыми-техникалық әдебиетті аудару үшін құрылып жатқандығы белгілі. Сондықтан қазіргі кезде қазақ сөзінің «Тіл – қазына» атты деректер базасын терминологиялық деректерімен де қызмет етуін алдын ала қарастырудың да маңыздылығы күннен-күнге артуда.

Болашақта қазақ сөзінің **«Тіл – қазына»** атты деректер базасы адам-компьютер қарым-қатынасының барлық түрін терминологиялық ақпаратпен қамтамасыз ететін болады. Мұнда тәжірибе жүзінде қолдану үшін ғылыми-техникалық терминдерді семантикалаудың компьютерлік жүйесін жасау және енгізу міндеті алдыңғы орынға шығады. Міне, осы аталғандар қазақ сөзінің **«Тіл – қазына»** атты деректер базасын құру мақсатындағы күн тәртібіне қойылатын ең маңызды мәселелер деуге болады.

Қазіргі кезде халықаралық қатынас нығайып, көптеген дербес ұлттық мемлекеттердің пайда болуы және бұрынғы отарлықта қалған халықтардың мәдениетінің көтерілу сипаты, қайтадан лексикографияның тіларалық қарым-қатынасқа және тілдерді оқытуға байланысты қызметтерін бірінші орынға ауыстыруына жағдай жасауда. Әлемдік өндірісті интеграциялау мәселесінде әртүрлі түсініктерді бірізге түсіру ісі де аса қажет болуда. Сондықтан егемен Қазақстанның халықаралық деңгейге шығуына және бірқатар жетекші елдермен қазіргі таңдағы ғылым мен технологияның әртүрлі салаларында ғылыми және өндірістік ақпараттармен алмасудың едәуір ұлғаюына байланысты неғұрлым кең ауқымды қамтитын көптілді сөздіктер шығару қажеттілігі де туындауда. Жақын болашақтағы лексикографиялық зерттеулердің негізінде оқу сөздіктерінің әр ұлттық тілдерге, соның ішінде қазақ тіліне қатысты да практикалық маңызы зор жүйесі құрылуы мүмкін.

Қазақстан Республикасының Тәуелсіздік алуына байланысты бүгінгі таңда ана тіліміздің өз мәртебесіне сай қоғамның барлық саласында қолданылу барысы жан-жақты жүріп жатыр деуге негіз бар. Осыған орай, толық сипаттағы болашақ қазақ сөзінің «Тіл – қазына» деректер базасының ішінде неше түрлі мәтіндер, сөзтұлғалардың статистикасы, сөзнұсқағыштардың толық тізімдері, сөздерге телулі цитаталар, жиілікті және кері-әліпбилі сөздіктер көрініс табатыны сөзсіз. Аталған тақырып өзегіне қатысты барлық еңбектер, ғылыми зерттеу жұмыстары бірізге, бір жүйеге түсіп, зор жетістіктерге жетеді деп білеміз.

Әдебиет

1. *Жұбанов А.Қ.* Қазақ тілінің компьютерлік қорын (базасын) құрастыру жайлы // Жұбанов А.Қ. Компьютерлік лингвистикаға кіріспе: Оқу құралы. – Алматы: Қазақ университеті, 2007. – 204 б. (189-199-бб.).

2. *Караулов Ю.Н.* Методология лингвистических исследований и Машинный фонд русского языка // Машинный фонд русского языка: идеи и суждения. –М.: Наука, 1986.

3. *Лейчик В.М.* Прикладное терминоведение и его направления. В кн.: Прикладное языкознание. Учебник. Отв. ред. А.С.Герд. Изд. Санкт-Петер. универ, 1996. –С. 276-286.

4. *Герд А.С.* Основы научно-технической лексикографии: Как работать над терминологическим словарем. –Л., 1986. 73 с.

5. *Караулов Ю.Н.* Методология лингвистических исследований и Машинный фонд русского языка // Машинный фонд русского языка: идеи и суждения. –М.: Наука, 1986.

6. *Реммель М.Н.* Выступление на Всесоюзной научной конференции по проблемам создания Машинного фонда данных для автоматизированной системы лексикографических исследований (Москва, 21-23 февраля 1983 г.) // Машинный фонд русского языка: идеи и суждения. –М.: Наука, 1986.

7. *Беляева Л.Н.* Машинный фонд русского языка для обучения языкам // Инженерная лингвистика и преподавание иностранных языков с помощью ТСО. –Л.: ЛГПИ им. А.И. Герцена, 1981.

8. *Караулов Ю.Н.* Методология лингвистических исследований и Машинный фонд русского языка // Машинный фонд русского языка: идеи и суждения. –М.: Наука, 1986.

9. *Шайкевич А.Я.* Проблемы терминологической лексикографии // Обзор. информ. Сер. 1. Теория и практика науч.-техн. пер. –М.: ВЦП, 1983. Вып. 8.

10. *Беляева Л.Н.* Машинный фонд русского языка для обучения языкам // Инженерная лингвистика и преподавание иностранных языков с помощью ТСО. –Л.: ЛГПИ им. А.И. Герцена, 1981.

11. *Беляева Л.Н., Герд А.С., Убин И.И.* Автоматизация в лексикографии. «Прикладное языкознание» учебник. Отв. ред. А.С. Герд. –СПб.: Изд-во С.Петерб. университета, 1996. –С. 318-333.

12. *Гончаренко В.В., Добрускина Э.М., Котельникова Н.М. и др.* Автоматическая переработка текста. –Кишнев, 1978.

ҰЛТТЫҚ ФИЛОЛОГИЯМЫЗДЫҢ ТАРИХЫНДА АЛҒАШ ПРОФЕССОР АТАҒЫ БЕРІЛГЕН ҚҰДАЙБЕРГЕН ҚУАНҰЛЫ ЖҰБАНОВТЫҢ ТАҒДЫРЫ ЖАЙЫНДА

ҚР БҒМ ҒК А.Байтұрсынұлы атындағы Тіл білімі институты. Көрнекті фольклортанушы ғалым, филология ғылымдарының докторы, профессор Есет Жұбановтың туғанына 80 жыл толуына арналған «Эпос тілін зерттеудің өзекті мәселелері» атты халықаралық дөңгелек үтел материалдары. Алматы-2011, 49-58-бб.

Біздің еліміздегі 30-жылдардың соңында үрей туғызған қуғын-сүргін

туралы аз жазылмаса да, өкінішке қарай, бұл тақырыпқа нүкте қою мүмкін емес. Ол жылдары түркі тілдерін, КСРО құрамына енген түркі халықтарының әдебиетін, тарихын, мәдениетін зерттеуге жеткілікті түрде назар аударылмады.

Осы дүрбелең кезінде 1937 жылы тамыз-қараша айларында КСРО ҒА Шығыстану институтының көптеген қызметкерлері жапа шекті. Олар негізінен Қиыр Шығыс: Жапония, Қытай, Корея, Моңғолия, Маньчжурия, Тибет елдері бойынша мамандар еді [1].

Қазақстандағы ұлттық зиялы қауым арасындағы зәбір көргендер одан да айтарлықтай көп болды. Қуғын-сүргіннің екі толқыны болды: 1929-1931 және 1937-1938 жж.

50-жылдары КСРО-ның көптеген республикаларының қуғын-сүргін көрген ғалымдары мен мәдениет қайраткерлерінің көпшілігі ақталғанымен, Қазақстанда негізінен қаза болған коммунистер ғана ақталып жатты. «Алаш» партиясы мен «Алашорда» мемлекетінің қызметіне қатысқан ірі мәдениет қайраткерлері 1988-1992 жылдары ғана ақталды. Қазіргі кезде Қазақстанда солардың ғана есімдері барынша құрметпен аталады.

Кеңес үкіметін мойындамаған «Алаш» партиясының мүшелері 1917 жылы Орынбор қаласында өткен құрылтайда Ә. Бөкейханов басқарған «Алашорда» атты өз мемлекетін құрды. Оның құрамына Ахмет Байтұрсынұлы да енді. 1929 жылы 12 маусымда А.Байтұрсынұлы тұтқынға алынып, РСФСР ҚК 58-7, 58-11 және 59-3 баптарына сәйкес қылмыстық жауапкершілікке тартылып, контрреволюциялық ұлттық «Алаш» ұйымын және «Алашорда» мемлекетінің негізін қалаушылардың және басшыларының бірі ретінде айыпталды. Үкімнің орындалуы туралы құжатқа (105-п.) сәйкес Ахмет Байтұрсынұлы 1937 жылдың 8 желтоқсанында атылды [1].

Ахмет Байтұрсынұлы ақталғаннан кейін оның есімі Қазақстанда, әсіресе 1991 жылы тәуелсіздік жарияланғаннан кейін ресми түрде жиі айтыла бастады. Ғалымның аты Алматының орталық көшелерінің біріне, Қазақстан Республикасының Ұлттық академиясының Тіл білімі институтына берілді.

Біз бұл мақаламызда Ахмет Байтұрсынұлынан кейінгі қазақ тіл білімінің көрнекті ғалымдарының бірі – Құдайберген Қуанұлы Жұбановтың тағдыры туралы қысқаша баяндауды жөн көріп отырмыз [1].

Қазақ филологиясы бойынша тұңғыш профессор, белгілі тілші-ғалым Құдайберген Қуанұлы Жұбановтың да әділетсіздік пен қатал тағдырдың құрбаны болғаны баршаға мәлім. Қазақ кеңес тіл білімінің пайда болуы мен қалыптасуын сөз ете отырып, біз отандық ғылымның дамуына жанын салған жандарды үлкен құрметпен еске аламыз. Осындай жандардың бірі 20-ыншы және 30-ыншы жылдары қазақ филологиясы бойынша жемісті еңбек еткен қазақ кеңес лингвистикасының негізін қалаушылардың бірі саналатын профессор Құдайберген Қуанұлы Жұбанов болды. Ол қазақ филологиясы ғылымының тарихында «профессор» атағын алған алғашқы ғалым еді [2; 3].

«Қазақ КСР тарихы. Социализм дәуірі» атты еңбектің соңғы басылымында 30-жылдардағы республикадағы ғылым мен білім беру, ағарту ісінің жай-күйіне егжей-тегжейлі сипаттама беріледі. Мына шағын азатжолға

назар аударалық: «В годы второй пятилетки первый педагогический вуз – Казахский педагогический институт имени Абая – превратился в крупное учебное заведение, важнейший культурный очаг... Здесь разрабатывались важные проблемы литературоведения и языкознания, истории Казахской ССР. Профессор института Х.К. Жубанов проделал значительную работу в области разработки казахской орфографии и новой, соответствующей особенностям казахского языка, письменности. Перу Х.К. Жубанова принадлежат первые учебники и исследования по грамматике, диалектологии и истории казахского языка» [4].

Тарихи хроника көп нәрсені тәптіштеп айтпағанмен, осындай мәліметтің астарында өмірінің мәнін туған халқына қызмет ету деп ұққан ірі ғалымның тұлғасы тұрғандығы айқын.

Құдайберген Қуанұлы Жұбанов 1899 жылы Орал облысы (Қазіргі Ақтөбе болысы) Темір уезі Темір Орқаш болысы Құсуақтам елді мекенінде (Темир-Уркачевская волость) орташа шаруа отбасында дүниеге келеді.

Қ.Қ. Жұбанов 1914 жылы «Хусайния» медресесінде оқу үшін Орынбор қаласына барады. Мұнда ол дүнияуи ғылымдарға көңіл қойып, орыс тілін үйренеді. Осы жерде ана тілін оқыту әдістемесінің теориялық курсының негізін салушылардың бірі – Шамғали Харесұлы Сарыбаев сияқты достар табады.

1916 жылы Құдайберген Жұрын екі сыныптық орыс мектебінің төртінші сыныбын бітіреді, бір жылдан кейін, яғни 1918 жылы 5 және 6-сынып бағдарламаларын бір оқу жылында қатар өтіп, Ілек училищесін үздік бітіреді. Осымен Қ.Қ. Жұбановтың білім алуының бірінші кезеңі аяқталады. Ол оқуын «Оспан-ишан» мешітінде Құраннан бастап, орта білімді Ілек орыс училищесінің қабырғасында аяқтайды.

Қ.Қ. Жұбановтың бастапқы кезде алған білімі оның тілші-ғалым болып қалыптасуына негіз болғанын айта кетуіміз керек. Өйткені ол осы жылдары араб және парсы, сондай-ақ орыс тілі сияқты құрылымы өзге тілдерді меңгереді. Он сегіз жасында-ақ Қ.Қ. Жұбанов түркі, араб және парсы тілдерінің сөйлеу, сондай-ақ жазба тілдерін меңгеріп шығады.

Қ.Қ. Жұбанов бала кезінен әдебиетке қызығып, кейбір шығыс тілдерін білгендіктен, көптеген араб-парсы әдебиетін, атап айтқанда, Фирдоусидің «Шахнамасының», Сағдидің «Гүлістанының», Низамидің «Ләйлі мен Мәжнүнінің» және т.б. классикалық туындылардың түпнұсқасын оқиды. Классикалық Шығыстың тілі мен әдебиетінен алған білімі оған кейіннен Абай атындағы Қазақ педагогикалық институты қазақ тілі кафедрасының профессоры болып, Орта Азия мен Қазақстан халқының мәдениеті мен өнеріне арналған әртүрлі мәселелерге қатысты көптеген іс-шараларға белсенді қатысқан кезінде көп пайдасын тигізді.

1925 жылы Құдайберген Ақтөбе губОНО-сында инспектор-әдіскер болып жұмыс істейді. Ол сол кездің өзінде бірден үш шет тілін – неміс, ағылшын және француз тілдерін үйренуге бар ынтасын салады. Алғашқыда оған қала мектептерінің оқытушылары мен жеке адамдар сабақ берсе, кейін өз бетінше оқып, айтарлықтай табысқа жетеді. Осы кездері академик И.П.

Павловтың, В.М. Бехтеревтің және басқа да көрнекті орыс ғалымдарының еңбектері бойынша психология, педагогика, логика және рефлексология пәндерін үлкен қызығушылықпен оқиды, Мәскеу, Қазан, Баку, Ташкент, Ленинград сияқты қалалардан Е.Д. Поливанов, В.А. Богородицкий, И.И. Мещанинов, Агамалы-оглы және т.б. көптеген ғалымдарымен хат алысады.

1928 жылдың көктемінде Қ.Қ. Жұбановты Қазақстан Республикасының сол кездегі астанасы Қызылорда қаласына шақырады, онда ол Халық Ағарту комиссариатының (Наркомпрос) инспекторы болып жұмыс істейді. 1929 жылы қаңтар айында ғалымды Ленинград Шығыстың тірі тілдері институтына түркологиялық семинар бойынша аспирантураға іссапарға жібереді, бір жылдан кейін сол жақтан КСРО Ғылым академиясының Яфетология институтына аспирантураға ауысады.

Қ.Қ. Жұбанов өз бетінше жалпы тіл білімі және түркітану саласында ғылыми-зерттеу жұмыстарын жүргізеді. Ол жаңа латын әліпбиіне өтуге арналған іс-шараларға белсене араласады, 1929 жылы маусым айында қазақ орфографиясының негізгі қағидаттарын әзірлеу мәселесіне қатысты Қызылорда қаласында шақырылған ғылыми-орфографиялық конференцияда сөз сөйлейді.

1930 жылы Қ.Қ. Жұбанов Алматы қаласында өткен Қазақстанның Бірінші өңіртану құрылтайында «Смыслы бессмысленных слов» деген тақырыпта баяндама жасайды. Бұл баяндама оның бірнеше түркі тілін салыстыра зерттеген тиянақты ғылыми ізденістерінің нәтижесі ретінде көрінді. Ал «Шаруашылық одағайлары» («Скотоводческие междометия казахов») атты баяндамасы үй жануарларына қаратылып айтылатын қазақ одағайларын талдауға арналған.

Кейбір деректерге сүйенсек, 1937 жылы Қ.Қ. Жұбановтың «Материалы к научной грамматике» (том I, Фонетика), «Еще раз о глаголе *де»* және т.б. бірнеше жұмысы жарық көруі тиіс болатын. Қазіргі кезде олардың тағдыры белгісіз.

Қ.Қ. Жұбанов түркі және басқа тілдердің білгірі ғана емес, қабілеті ерекше публицист болған. Оның «Күй» [5] атты тамаша туындысы тақырыптың сонылығымен және деректік материалдардың молдығымен ерекшеленеді.

Жалпы алғанда Қ.Қ. Жұбанов осы жұмыста филологтарға ғана емес, Абдул-Хадир Мураги, Рауф Екта, Әл Фараби, Кизеветтер, Беляев, Успенский, проф. Фитрат, проф. Оганезашвили және т.б. авторлардың музыкатану бағытындағы жекелеген еңбектерін келтіре отырып, түркі тілдерінің тарихи-лингвистикалық деректерін жүйелеу мен талдау арқылы халық музыкасы жанрларының пайда болу ретін нақтылап, қазақ музыкатанушылары мен өнер тарихшыларына жол сілтеді.

Қ.Қ. Жұбановтың қазақ жазба әдебиетінің классигі Абайдың қайтыс болғанына 30 жыл толуына орай ақынның тілі туралы жазылған мақаласы оның публицистикалық еңбектерінің ішіндегі танымдық мәні жағынан қызықты да мазмұнды шығармаларының бірі болып табылады. Бұл мақала

қазақ әдебиеттану ғылымына қосылған аса құнды үлес десек артық айтқандық болмас. Қ.Қ. Жұбанов Абай өлеңдері құрылысының төлтумалығын, оның халық тілін қолданудағы жаңашылдығы мен шеберлігін алғашқылардың бірі болып дәлелдеді. Жалпы әлем мәдениетінің, оның ішінде классикалық Шығыс мәдениеті қазынасының тамаша білгір ретінде Қ.Қ. Жұбанов осы мақалада Абай Құнанбайұлының қазақ халқы мәдениетінің тарихындағы орны мен мәнін дұрыс анықтау үшін өткенге еркін «саяхат» (экскурс) жасайды.

Қ.Қ. Жұбанов көрнекті түркітанушы зерттеуші ретінде лайықты түрде саналады. Ленинград Шығыс институтының, содан кейін Яфетидология институтының (кейіннен КСРО ҒА Тіл білімі институты) аспиранты бола жүріп, өзін шығыстану саласында зерделі бақылаушы, кең ауқымды тілші-зерттеуші ретінде танытты. Қ.Қ. Жұбановтың дарыны мен қабілеті өзі бірнеше жыл қазақ тілі кафедрасының меңгерушісі болып қызмет еткен Абай атындағы Қазақ педагогикалық институтында сабақ берген жылдары ерекше айқын көрінеді. Ол көне түркі жазулары тілін мүлтіксіз меңгеріп, тыңдаушыларды руна (Орхон-Енисей) жазулары мәселелері, олардың лексика-грамматикалық, фонетикалық ерекшеліктері туралы білім нәрімен сусындатты. Қ.Қ. Жұбанов өзінің педагогикалық қызметін қазақ және басқа да түркі тілдерінің ғылыми мәселелерімен шебер ұштастыра білді. Ғалымның қаламынан туған ғылыми еңбектер қазіргі уақытта да өзектілігін жоғалтпады. Оның еңбектеріне салыстырмалы-тарихи зерттеу әдісі тән.

Тілдік құбылыстардың тарихи негізін түсінбей тұрып шынайы ғылыми талдау жасау мүмкін емес. Қ.Қ. Жұбановтың сол кездері «сәнді» болған лингвистикалық бағыт – яфетидология мәселелеріне қызығушылық танытуы оның артында қалдырған ғылыми мұрасының құндылығын түсіре алмайды.

Көмекші және күрделі етістіктерге арналған жұмысында [6] автор өзін ана тілінің лексика-грамматикалық тетігінің білгірі ретінде танытты. Қ.Қ. Жұбановтың аталған еңбегі тіл тарихшыларын сөйлем мүшелерінің ұйымдасу мәселелерін зерттеумен ғана емес, бірнеше түркі тілдерінің грамматикалық және лексикалық формаларының этимологиясын анықтаумен қызықты. Бұл еңбек түркі тілдерінің деректерін зерттеуде салыстырмалы-тарихи әдісті қолданудың тиімділігін көрсетті.

Қ.Қ. Жұбанов қаламынан туған қазақ орта мектептерінің 5-6-сыныптарына арналған алғашқы оқулықтардың бірі саналатын қазақ тілінің грамматикасын ерекше атап өтуіміз керек [7]. Бұл оқулықта автор қазақ тілінің сөз құрамының негізгі мәселелері мен фонетикалық құбылыстарын қатаң жүйеге салып берді. Бұл грамматика орта мектептерге арналған қазақ тілі оқулықтарын жазуда үлгі саналады.

Ғалым қазақ фонемаларының сандық және сапалық құрамы туралы өз ойын ортаға салады. Проф. Қ.Қ. Жұбанов осы саладағы өз зерттеулерінде «буындардың кірігуі мен алмасуы (слияние и смещение слогов), екпін, әсіресе, үндестік заңдары сияқты көптеген фонетикалық құбылыстар сөз бөліктерінің өзара мықты байланысуына қызмет етеді», – деген ғылыми тұжырымға келеді. Ғалым әліпби, орфография және қолданбалы тіл білімінің

басқа да салаларына арналған кейінгі еңбектерін қазақ фонологиясының ережелеріне негіздеген.

Қ.Қ. Жұбанов 1936 жылы грамматикасымен қатар орта мектептердің 5-7-сыныптарына арналған қазақ тілі бағдарламасын дайындады. Бұл бағдарлама қазақ тілінің ғылыми зерттеу пәніне айналғандығын растайтын алғашқы құжаттардың бірі болып табылады. Онда Қ.Қ. Жұбановтың құрмалас сөйлем синтаксисіне қатысты көзқарастары айқын көрінеді. Қ.Қ. Жұбановтың бұл оқулығы сол кездері республикамыздың оқытушылар жұртшылығынан жоғары бағасын алды. Қ.Қ. Жұбановтың бұл жобасы туралы белгілі кеңес түркітанушысы С.Е. Малов та жоғары пікірде болды. Ол 1935 жылы мәдени құрылыс жөніндегі құрылтайдағы сөзінде «проект профессора Х.К. Жубанова чрезвычайно научный, принципиально выдержанный», – деп атап көрсетті [8].

Бұл айтылғандардың бәрі Қ.Қ. Жұбановтың қазақ мектептері тарихында көрнекті педагог ретінде лайықты орын алатындығын дәлелдейді. Алайда жеке басқа табынушылық кезеңінде Қ.Қ. Жұбановтың қазақ тілі грамматикасы әділетсіз сыналды: грамматикалық ережелерден «зиянкестік» пиғыл іздеген мақалалар пайда болды.

Қ.Қ. Жұбановтың тұтас мақалалар топтамасы практикалық қазақ тілінің орфографиясы, терминологиясы, қазақ тілін оқыту әдістемесі және т.б. мәселелерге арналды. Тілдік құбылыстарды зерттеудің төлтума сипаты, филолог мамандар дайындаудағы қызметі арқылы Қ.Қ. Жұбанов қазақ кеңес тіл білімінің негізін салушылардың бір болды. Профессор Құдайберген Қуанұлы Жұбановтың есімі өз кезінде қазақ зиялы қауымының, әсіресе білім беру-ағарту саласы қызметкерлерінің арасында танымал болды. Оны академиктер И.И. Мещанинов, А.Н. Самойлович, В.В. Бартольд, С.Е. Малов, К.К. Юдахин, А.К. Боровков және т.б. сияқты орыс ғалымдары тамаша лингвист-ғалым ретінде танып, еңбектерін жоғары бағалаған.

Ғылыми-педагогикалық қызметімен қатар Қ.Қ. Жұбанов қоғамдық өмірге де белсене араласты. Фирдоуси, Абай сияқты көрнекті қайраткерлерге арналған ғылыми конференциялар ұйымдастыруда бастама көтеріп, осы конференцияларда мазмұнды баяндамалар жасады. 1935 жылы оны Қазақстандағы мәдени құрылыс қызметкерлері құрылтайының ұйымдастыру комитетінің төрағасы етіп, содан кейін XVI Бүкілресейлік Кеңестер құрылтайына Қазақстан атынан делегат етіп сайлады.

1937 жылы қазақ арасынан шыққан кәсіби тілші ғалымдар өте аз болды әрі олардың барлығын дерлік (өз қызметтерін енді ғана бастап келе жатқан ең жас зерттеушілерді қоспағанда) құртып жіберді. Ал қазақ тіл білімін қайта жандандыру қажет болған кезде кәсіби тілші ғалымдардан ешкім қалмаған еді. Сталиндік жазалаушы мәшине дәл әрі мүлтіксіз жұмыс істеді. Шығармашылық күші мен қуаты толысқан, дарыны пісіп жетілген ғалым, зерттеуші, отыз сегіз жастағы профессор Құдайберген Жұбановты ұлтшыл ретінде айыптады. 1936 жылы жарық көрген қазақ орта мектептеріне арналған төлтума грамматика осы айыптауға негіз болды. Қ.Қ. Жұбановтың оқулығында 1937 жылдың басында кенеттен «халық жауы» аталған С.

Сейфуллин, Б. Майлин, І. Жансүгіров сынды ақын-жазушылардың шығармаларынан материалдар келтірілген болатын. Қазақ жазуының орфографиялық нормасын қалыптастыру мәселесінде Қ.Қ. Жұбанов сол кезде контрреволюционер ретінде саналған Ахмет Байтұрсынұлы принциптерінің бірін қолдады. Осы және басқа да жағдайлар Жұбановты ұлтшыл-фашист ретінде айыптауға «айғақ» болды.

Ғалымның темір тордың ар жағындағы өмірінің соңғы күндеріне қатысты кейбір деректерді келтіре кетсек.

Қ.Қ. Жұбановтың табандылығы мен алған бетінен қайтпайтындығы жөніндегі деректердің бірі Қазақстан Республикасы Ұлттық қауіпсіздік комитетінің (ҰҚК) мұрағатында сақтаулы «РСФСР ҚК 58-бабының 2, 8, 9 және 11-тармақтарына сәйкес айыптау бойынша №11020 істе» келтірілген мәліметтер болып табылады. Онда «1937 ж. 19 қарашада басталған» және «1937 ж. 23 желтоқсанда аяқталды» деп көрсетілген «№9555387 мұрағаттық-тергеу ісі» бар [9, 102-107].

«Тергеу хаттамасында» профессор Қ.Қ. Жұбановтың мына сөздерінің астарында терең ой жатқандығын аңғарамыз: «Наши руководители кричали о вражеских кознях баев, о сопротивлении байства социалистическому переустройству Казахстана, науськивали на байство и без того ретивых активистов. Забывали при этом о том, что байство – цвет и сила казахского народа, что не беднота и голытьба кормили страну, а те же «презренные» баи давали государству скот, мясо, шерсть, кожу и хлеб. А теперь когда конфисковали байство, пустили по ветру неисчислимые богатства нашей страны, развалили веками сложившееся хозяйство Казахстана, теперь оставшись «у разбитого корыта», мы начали кричать о перегибах, об их последствиях, о необходимости их исправления и т.п.».

Ғалымның «голошекин» кезеңіндегі асыра сілтеушілікке қатысты ойлары, біздіңше, өте батыл әрі қисынды айтылған. Оқырманның ғалымның мына ойларымен келіспеуі мүмкін емес: «Вот и мы говорим, что были перегибы при Голощекине. Что значит эти перегибы? – Это значит – смерть значительной части казахского народа, и только казахского народа. Это лишение его единственного богатства – скота, посредством непосильных поборов, так называемой скотозаготовки. В результате была массовая гибель казахского народа. Русские в Казахстане не голодали и не умирали...».

Қ. Жұбанов бүкіл қазақ халқының тауқыметті тағдырының себебін іздегендігі мына сөздерінен көрінеді: «...Сейчас исправляя перегибы, опять судим тех же казахских работников. Получается что и так, и эдак погибают казахские работники. При таком положении вымирание казахского народа неизбежно... За растранжирование денег – людей судят, а почему же растранжирование нации, целого народа не считается преступлением? Конечно, это дело не одного Голощекина. Все это происходит на глазах Советской власти, тем не менее она не обращает на это никакого внимания. Это потому, что она строит социализм и социализм она будет строить, если даже неизбежно жертвовать отдельными народами. Таким образом, казахский народ и явился жертвой социализма. Наша задача должна

заключаться в том, чтобы защищать казахский народ от вымирания, от того, чтобы он не стал жертвой социализма...».

Біздің ойымызша, Қ.Қ. Жұбановтың жазаланар (атылар) алдында айтқан мына сөзі ешкімді де бейжай қалдырмасы анық: «... при Советской власти, и наша культура не может быть национальной, а неизбежно будет руссифицирована. Если мы хотим, чтобы казахский народ не был жертвой социализма, то надо освободить его от этого социализма; а этого мы можем добиться только путем создания независимого буржуазно-демократического государства, чего бы нам это не стоило».

Сонымен қазақ лингвистикасының негізін қалаушылардың бірі, қазақтың тұңғыш профессоры Қ.Қ. Жұбановтың 70 жыл бұрын айтқан идеялары мен пікірлері қазақ халқының қазіргі жағдайымен үйлеседі.

Жеке басқа табыну кезінде Қ.Қ. Жұбанов жасына байланысты «Алашорда» қызметіне қатыса алмады, оны 1937 жылы тергеушілер де ғалымның мойнына таңа алмады. Оның ісінде (201-п.) мынадай жолдар кездеседі: «Еще с 1917 года Джубанов принимает самое активное участие в издательстве казахского националистического журнала «Тез» и в последующем, состоя членом националистической группировки, издавал националистический журнал «Кызгалдак»» [1].

Кейін белгілі болғандай, ғалым 58-баптың 2, 8, 9 және 11-тармақтары бойынша айыпталған. Қ.Қ. Жұбановты алдымен бірінші бап (58-2) бойынша жапон жансызы ретінде айыптады. Бұл айыптауға оның жапон тіліне қызығушылығы себеп болған сияқты. Кейін бұл бап (жансыздық) алынып тасталды, алайда қалған үшеуі, әсіресе 58-8 бабы (лаңкестік) өлім жазасына кесуге жеткілікті еді. Осылайша Құдайберген Қуанұлы Жұбанов 1938 жылы 25 ақпанда атылады [1].

Ғалым қаза болғаннан кейін Қ.Қ. Жұбановтың отбасында зайыбы Раушан мен оның алты баласы (ең үлкені бар-жоғы 12 жаста, ең кішісі 3 айлық) қалды, олар кейіннен жоғары білім алып, филология саласында әкелерінің ісін жалғастырып келеді.

Жұбанов ісінде оның «Алашордаға» қатысы жоқтығы оны өлім жазасынан арашалап қала алмады, алайда ақталуға септігін тигізді.

Қ.Қ. Жұбановтың күнәсіздігіне көз жеткізу үшін оның ғылыми еңбектері туралы филология ғылымдарының кандидаты Махмудов и Сайкиевтен, филология ғылымдарының докторы М. Балақаевтан, ҚазССР Ғылым академиясы өнертану секторының меңгерушісі Ерзаковичтің пікірлері алынған: «Все эти ученые отзываются об опубликованных трудах Джубанова положительно и заявляют, что они имеют определенную научную ценность и не содержат никаких антипартийных или антисоветских идей или положений» (л. 211) [1].

Әрі қарай мынадай қорытынды шығарылады: «Настоящее дело... внести на рассмотрение Военной коллегии Верховного суда СССР на предмет отмены приговора и прекращения дела в отношении Джубанова Худайбергена за отсутствием в его действиях состава преступления» (212-п.) [1]. 03.04.1957 ж. Қорытындыны КСРО Бас прокурорының орынбасары

бекітеді, ал 03.10.1957 ж. КСРО Жоғарғы сотының Әскери коллегиясы 1938 жылғы үкімді жоққа шығарады (213-п.). Нәтижесінде Қ.Қ. Жұбановтың еңбектері ғылыми айналымға А. Байтұрсынұлының еңбектерінен бұрын түсіп, ғылыми еңбектерінің бастылары 1966 жылы қайта басылып шықты [1].

Қазақ лингвистикасының негізін салушы, көрнекті түркітанушы-ғалым Құдайберген Жұбановтың есімі 1957 жылға дейін қазақ халқының жадынан өшіріліп тасталған еді. Ұшқыр ойлы, гуманитарлық пәндер саласында шынайы энциклопедиялық білімді біртуар ғалымнан мезгілсіз айырылу 30-жылдары әлі қалыптасып үлгермеген қазақ тіл ғылымының даму үдерісіне айтарлықтай әсер етті. Алайда оның бар-жоғы 20 жыл ішінде жасаған дүниелерін кейінгі ұрпақ лайықты бағалап отыр. Профессордың мәңгі өмірі оның құнды еңбектерімен, жас тілші ғалымдар буынының ізденістерімен жалғасып келеді.

Қазіргі кезде Жұбанов үрдісі ғылымда жалғасын табуда. Оның ғылыми еңбектері қайта басылып шығып жатыр. Республикамыздың ірі ғылыми мекемелері мен жоғарғы оқу орындарында мерейтойлық күндері аталып, қазақ филологиялық ғылымының тарихында алғаш профессор атағы берілген ғалымға арналған ғылыми-теориялық конференциялар өткізілуде. Ғалымның туған жері – Ақтөбе қаласында Мемлекеттік университетке Құдайберген Жұбанов есімі беріліп, ескерткіш орнатылды және университет алдындағы даңғыл ағайынды проф. Құдайберген мен оның туған інісі академик Ахметтің атымен – «Ағайынды Жұбановтар» аталды.

Қазақ тіл білімінің негізгі проблемаларына іргелі лингвистикалық зерттеулері мен осындай зерттеулердің нәтижелерін қолданысқа енгізудегі ғылыми-практикалық қызметі Қ.Қ. Жұбановты оның өз дәуіріндегі ғылымның көрнекті өкілі ретінде танытады.

Әдебиет

1. *Ашнин Ф.Д., Алпатов В.М., Насилов Д.М.* Репрессированная тюркология. – М. : Вост. лит., 2002. – 296 с.

2. *Ауэзов М.О., Кенесбаев С.К.* Основоположник казахской лингвистики// Худайберген Жубанов и казахское советское языкознание. Алма-Ата: Наука, 1990, стр. 21-25.

3. *Кенесбаев С.К.* Профессор Худайберген Куанович Жубанов // *Х.К. Жубанов.* Исследования по казахскому языку. - Алма-Ата: «Наука», 1936. – 361 с.

4. «История Казахской ССР. Эпоха социализма». Алма-Ата, 1963, стр. 529-531.

5. *Жубанов Х.К.* К вопросу о возникновении казахского музыкального жанра – кюй. Историко-лингвистический очерк. Кзыл-Орда, 1936 (на каз. яз.).

6. *Жубанов Х.К.* Исследования по казахскому языку. Вып. 2. Заметки о вспомогательных и сложных глаголах. Алма-Ата, 1936.

7. *Проф. Қ. Жұбанов.* Қазақ тілінің грамматикасы, I бөлім. Сөздің жалпы құрылысы (жалпы морфология). Орта мектептің V класына арналған. 1936, Қазақстан баспасы, 74 б.

8. Стенограмма выступления С.Е. Малова на съезде по культурному строительству в Казахстане. (См. ЦГА КазССР, ф. 81, оп. 3, д.1086, лл. 10-12).

9. *Жұбанов Е.Қ., Оралбай Ә.* Профессор Құдайберген Жұбановтың саяси памфлеттері // Жұбановтану жылнамасынан. – Ақтөбе, 2004. 186 б.

ТҮРКІ ТІЛДЕРІНІҢ КӨПАСПЕКТІЛІ КОМПЬЮТЕРЛІК ҚОРЫН ҚҰРУ ҚАЖЕТ

ҚР БҒМ ҒК А.Байтұрсынұлы атындағы Тіл білімі Институтының 50 жылдығына арналған «Қазақ тіл білімі: тарих, тәжірибе, болашақ» атты халықаралық ғылыми-теориялық конференция материалдары. – Алматы, 2011. – Б. 327-335.

Электронды есептеу машиналарының (ЭЕМ) көмегімен ұлттық тілдерді зерттеу мен одан туындайтын ақпараттарды автоматты түрде танып-білудің негізінде ХХ ғасырдың 80-жылдарынан бастап тілдердің машиналық қорын (базасын) жасау жұмысы жайлы ой-пікірлер жүзеге аса бастағаны белгілі. Осымен байланысты 1983 жылы Мәскеуде шақырылған Бүкілодақтық ғылыми конференцияның күн тәртібіне орыс тілінің машиналық қорын жасау мәселесі алғаш рет қойылды. Бұл ғылыми мәжілісте ЭЕМ-ның жадына түсірілетін тілдік қордың құрамы мен құрылымы қандай болуы тиіс екендігі, бірінші кезекте мұндай қор қандай міндеттер атқаратыны, тіл көздерінің ауқымы қандай болмақтығы жайлы және т.б. көптеген мәселелер қаралған болатын. Сонымен бірге электрондық тіл қорын дәйекті түрде тұтыну үшін қандай ұйымдық және техникалық шаралардың атқарылу қажеттігі ортаға салынып, жан-жақты талқыланды.

Аталған ғылыми конференцияның шешімдерінің орындалу жағдайын қадағалау мен жаңадан туындайтын мәселелерді шешу қажеттігінен 1987 жылы Мәскеуде ашылған Бүкілодақтық екінші ғылыми конференцияда осы саладағы ғалымдар тағы да бас қосты. Бұл ғылыми форумға Қазақстанның және басқа да ұлттық республикалардың ғалымдары қатысты.

Мәжілісте өткен төрт жыл аралығында орындалған жұмыстардың нәтижелерінің сөз болуымен қатар, басқа да ұлт тілдерінің машиналық қорларын жасау мүмкіншіліктеріне байланысты соны мәселелер қарастырылды. Ерекше атап кетерлік жәйт – осы ғылыми жиында орыс, украин, грузин, эстон тілдеріне қатысты баяндамалардың қатарында қазақ тілінің машиналық қорын жасау мәселесі де өз алдына тыңдалды.

Сөйтіп, орыс тілінің машиналық қорын жасау мәселесіне арналған екінші ғылыми конференцияның шешімдерінде қазақ мәтіндерін ЭЕМ

жадына енгізу және оны автоматты жолмен статистикалық тәсілдермен зерттеу бағытындағы қазақ ғалымдарының қол жеткізген нәтижелері ескеріліп, түркі тілдерінің машиналық қорын жасайтын координациялық орталық Қазақ ССР ҒА Тіл білімі институтының құрамынан ашылсын деген дербес қаулы қабылданғаны біз үшін үлкен абырой болды.

Көп ұзамай 1988 жылдың мамыр айында Мәскеуде Кеңес Одағы түркітанушы комитетінің XIV Пленумы ашылып, оның күн тәртібіне түркі тілдерінің машиналық қорын жасау мәселесі арнайы қойылды. Бұл ғылыми жиында негізгі баяндамашылар ретінде Одақ бойынша белгілі ғалымдар: Р.Г.Пиотровский, А.М.Щербак, В.Г.Гузев өз ой-пікірлерін ортаға салып, ондай қор жасаудың қажеттігімен қатар алда тұрған қиыншылықтарын да атап өтті [1]. Ал қазақ тілінің машиналық қорын жасау мүмкіндігі жайлы мәселеге байланысты қазақстандық ғалымдар баяндама жасады (Қ.Б.Бектаев, А.Қ.Жұбанов).

Тілдің машиналық қорының қажеттілігін алғаш дәлелдеп берген ғалым, информатика саласының белгілі маманы А.П.Ершов тілдің машиналық қорын құрастыру мәселесіне үлкен мән берген болатын [2]. Оған ғалымның 1978 жылы айтқан мына сөздері айғақ: «Орыс тілін сүйетін қоғамның (формалды қоғам) танымынша, орыс тіл білімі дегеніміз әзірге ұйықтап жатқан Гулливер тәрізді, сондықтан менің ойымша, оның оянатын және өзі туралы бар дауыспен жар салатын кезі енді жетті. Ал ЭЕМ бұл ретте таптырмайтын қоңыраулы сағат секілді» [3, 7-12-бб.]. Бұл айтылғандар қазақ тілінің қазіргі жағдайының компьютерлік лингвистика саласына қатысты әлі де өз күшін жоғалтпаған тәрізді. Ал А.П.Ершов сол кездің өзінде-ақ аса білгірлікпен болжап, болашақ жасалатын орыс тілінің машиналық қорын тіл жөніндегі білімдердің аса қуатты концентраты (қоспасы) болады деп сенген.

Орыс тілінің машиналық қорын құру ісіне алғашқы күнен басшылық етіп келе жатқан ғалым В.М. Андрющенконың сөзімен айтсақ: «Создание машинного фонда русского языка рассматривается не только как прикладная задача, проект предусматривает также проведение теоретических исследований и экспериментальных работ, он создает условия для развития лингвистики как подлинно экспериментальной науки, существенно повысит эффективность действующей справочной службы яыка», – дейді [4]. Міне, осы іспеттес, қазақ тілінің машиналық қорын құрастыру үшін де арнайы теориялық және тәжірибелік зерттеулер жүргізілуі қажет болады.

Мұндай формалды компьютерлік жүйе тілдің жаратылысына сәйкес аса көлемді болуы қажет. Сөйте тұра ондай тілдік модель «анатомиялық» тұрғыдан мүшеленген, ашық, бақылауға айқын, зерттеуге қолайлы әрі өзгермелі болуы тиіс деп ұйғарған жөн. Алайда, қазақ тілін модельдеуге қатысты біздің жетістіктерімізді үндіеуропа тіл білімінің статусымен және оның компьютерлік лингвистика бойынша алған нәтиже ауқымымен салыстырып қарайтын болсақ, біз өтпек жолымыздың әлі де бастауында тұрғанымыз айқын сезіледі.

Тілдік құбылыстардың біркелкі еместігі, бір қарағанда, компьютерді тіл білімі саласындағы зерттеу жұмыстарында пайдалануға кедергі келтіретін

сияқты. Егер біз компьютердің мүмкіндігін толық түрде ескере білсек, тіпті шешімі жоқ-ау деген көптеген мәселелердің де шешімі табылар ма еді. Көбінде, компьютер жадында орындалатын амалдар тілшінің орындайтын формалды іс-әрекетіне ұқсас және олар көрініс деңгейінде тіл мен сөйлеу аспектілерінің көптеген жақтарын қамти алатыны сөзсіз. Сондықтан бұл жердегі ең бастысы адам баласы өзінің ойлау процесін компьютерге бөлшектеп беріп, формалдау жолдарымен ұштастыра білуі қажет сияқты. Сонда ғана зерттеуші тілдік парадигмалық көріністерді, фразалық тармақтанудың заңдылықтарын ашуды және т.б. формалдануға қажетті жұмыстарды іске асырар ма еді.

Қазақ тіл біліміндегі зерттеу кезеңдеріне, әсіресе, лингвистикалық ақпараттармен жұмыс істеу: оларды жинақтау, іздеу, сақтау, шет тілдегі ақпараттарды қазақ тіліне аудару, ақпараттық библиографиялық іздеу, қазақ тіліндегі мәтіннен реферат алу және т.б. істерге логикалық роботтың (компьютердің) араласуы болашақта мемлекеттік тұрғыда қайта қаралып жатса, бұл мәселе өндірістік дәрежеге дейін көтерілуі де мүмкін.

Ғылым мен техниканың дамуы тіл қызметінің қоғамдағы коммуникативтік қолданыс аясын кеңейте түскені бәрімізге мәлім. Тіл өзінің адамдар арасындағы қарым-қатынастық қызметін атқаруымен бірге адам мен компьютердің қазақ тіліндегі диалогын да жүзеге асыруы қажет деп білеміз. Осы себептен жақын болашақта табиғи тіл және оның түрлері ақпарат таратудағы **«адам–компьютер–адам»** аралығындағы коммуникативтік жүйе негізін құрайды деп сенім білдіруге әбден болады. Әрине, аталған жүйеде – тіл, өзінің адам баласының ойлау қабілетіндегі үйреншікті ассоциативті-эвристикалық мүмкіншілігінен басқаша өзіне жат қалыпқа түсіп, компьютердің тізбектік-алгоритмдік субстратына ауысады. Тілдің бұл құбылысы қазақ тіл біліміндегі ізденістерге біршама методологиялық қиыншылықтар туғызуы мүмкін. Осы себептен, қазіргі тілдік зерттеулердің, біріншіден, қатаң логикалық теорияға бағындырылуын, екіншіден, объективті және толық болуын барынша талап ету қажет.

Тіл қызметінің жоғарыда сөз болған шарттары тіл білімінің «инженерлік лингвистика» деп аталатын жаңа саласында қарастырылады.

Инженерлік тіл білімі, негізінен, электрондық машина (компьютер) қажеттігі үшін тілдік белгілерді, тіл жүйесі мен оның атқаратын қызметін (функциясын) модельдеу мәселелерімен айналысады.

Тілдік объектілерді жасанды түрде қайта туындату үшін, тек қана жоғары дәрежедегі техникалық база болуы жеткіліксіз. Біздің ұйғаруымызша, қолданбалы есептер үшін, ең бастысы, күшті және объективті теория құрылуы қажет. Сондықтан да әрбір зерттеушінің тәжірибе жүзінде алған нәтижелері сол лингвистикалық теорияға баға беретіндей өлшемге айналуы керек.

Инженерлік тіл білімінің қажеттігін дәлелдейтін проблемалық бағыт-бағдарлардың ең негізгісі – мазмұн бірлігін автоматты түрде айқындап білу және табиғи тіл мен жасанды тілдерді семиотикалық жағынан барынша жақындастыру проблемалары. Сонымен қатар ашық жүйеге жататын табиғи

тіл және келісіммен алынған жабық семиотикалық жүйедегі қарым-қатынастарды зерттеу – инженерлік тіл білімінің методологиялық және технологиялық мәселелерінің ең негізгілері. Бұл мәселенің дұрыс шешім табуы – семиотика, информатика, компьютерлік бағдарлама жасау және математикалық машиналар теориясы, ойлау процестерін модельдеу салаларындағы ғылымдардың қарқындап дамуының бірден-бір шарты.

Инженерлік лингвистиканың тіл біліміндегі басқа да пәндер қатарынан тең дәрежеде орын алуы – көптеген тілшілер мен математиктердің, инженер-программистердің жүргізген тәжірибелерінің, теориялық ізденістерінің нәтижесі деп түсінген жөн. Мәселен, қазіргі қазақ тілінінің «Статистикалық лингвистика» және «Инженерлік лингвистика» салалары – физика-математика ғылымдарының кандидаты, филология ғылымдарының докторы, профессор Қ.Б.Бектаев есімімен және ол ғалымның шәкірттерінің көптеген ғылыми ізденістерімен тығыз байланысты. Сол сияқты, қазақ тілі грамматикасын формалды сипаттауға байланысты соңғы кездегі ғылыми ізденістерімен көзге түсіп жүрген техника ғылымдарының докторы, профессор А.А.Шәріпбаевты және т.б. атауға болар еді. Аталған зерттеу нәтижелері қазақ тілін инженерлік лингвистиканың әдіс-тәсілдерімен қарастыруда, мемлекеттік тілімізді оқытудың (үйретудің) автоматты түрдегі жүйесін құруда және қазақ тілінің формалды грамматикасының ғылыми негізін құру жолының бастамасы болатынына еш күмәніміз жоқ.

Классикалық лингвистика көбінде ақпараттық процестің қандай материалдық ортада жүріп жатқанына мән бере бермейді. Ал инженерлік тіл білімінде басқаша, яғни табиғи тілдің «адам–компьютер–адам» жүйесіндегі қызметінің ерекшеліктері стратегиялық және технологиялық ортаға тікелей қатысты. Сондықтан тілдік мәліметтерді компьютер арқылы қайта қарастыру жағдайында, адам баласының миы мен компьютер жадының құрылымды-функционалдық сипаттарын салыстыра отырып, оларға тән қасиеттерді ескерген жөн. Олай дейтініміз, тіл білімі саласындағы электрондық машиналар мен адам баласының мүмкіншіліктері жайлы көптеген тілшілер мен математиктер ғылыми әдебиеттер бетінде өз ойларын білдіріп жүрсе де, тіл білімі саласындағы компьютердің болашағы жайлы ойлары әлде де бірізді емес. Мысалы, біреулері ЭЕМ-ның жады мен жылдамдығын місе тұтпаса, ал басқалары – компьютер арқылы семантика-синтаксистік, тіпті стильдік ерекшеліктерге қатысты күрделі мәселелердің болатынын сөз етіп, қалай болғанда да компьютерді тілден алшақтатуға тырысатындар да жоқ емес.

Компьютерлік техниканың барынша дамыған кезінде оның жадының көлемі, жылдамдығы туралы сөз болмаса да, басқа ғылыми мәселелердің «кезек күттірмейтін» тұстары (мысалы, түрлі сөздіктер түзу) себеп болып, қазақ тілін зерттеуде компьютерлік тіл білімінің жолы ойлағанымыздай болмай, қарқыны бұрынғыдан да бәсеңсігенге ұқсайды.

Адам мүмкіншілігіне қарағанда, компьютер сырт кедергілерге қарамастан, жадындағы ақпаратты әрі тұрақты, әрі берік, әрі ұзақ мерзімге сақтайтыны және қажетті ақпаратты тез арада қағаз не монитор бетіне шығарып бере алатындығы көпшілікке мәлім. Ал мәтін ішіндегі ақпаратпен

(мысалы, мағына мен мазмұн) жұмыс жасауда адамның ми қызметі оның физиологиялық және психикалық қалпына байланысты болатын болса, мұндай қасиет компьютерге жат деуге болады.

Инженерлік тіл білімінің даму болашағын сөз еткенде, мысалы, жақын он жылдың ішіндегі зерттеудің стратегиясын анықтауда және үйлесімді технологияны таңдауда тек қана бүгінгі күнгі кибернетикалық техниканың және белгілі шеңбердегі тіл білімі мен нейрофизиологияның ғылыми жетістіктеріне сүйенген жеткіліксіз деп есептейміз. Бұл жерде кибернетикалық техниканың даму қарқынын ескере отырып, кибернетиканың, тіл білімінің және нейрофизиологияның методикалық жақтарын ескеретін кең шеңбердегі алдын болжай алатындай ғылыми көзқарас қажет.

Қазіргі кезде еуропа тілдеріндегі мәтіндік хабарламалардың грамматикалық, сөздік, фразеологиялық, кейбір лексика-грамматикалық және, тіпті, семантикалық деңгейлерін айқындайтын компьютерге арналып жасалған алгоритмдер негізіндегі бағдарламалары іске қосылған.

Осы айтылғандардың негізінде төмендегідей сұрақ туындауы мүмкін.

Электронды автоматтың (компьютердің) болашақтағы мәтінді түсіндіру мүмкіндігі қай деңгеймен шектелмек? Ол тек лексика-грамматикалық, лингвистикалық деңгей ме? Әлде компьютердің мәтін мазмұнын анықтаудағы мүмкіндігі маман аудармашының деңгейіне теңелуі мүмкін бе? Компьютерлік аударма проблемасына қатысты бұл сұрақты былай да қоюға болады: шетел тіліндегі белгілі бір сала бойынша алынған мәтінді әрі осы саланың маманы, әрі тәжірибелі аудармашы болып табылатын адамның аударуы мен компьютерлік аударманы салыстыруға бола ма? Әлде компьютерлік аударма еш уақытта жоғары сапалы болмай, оған тұтынушының амалсыз араласуы қажет пе?

Әрине, қазіргі кездегі компьютерлік аудармалар әлде де маман аудармашының жәрдемін қажет етеді. Сондықтан да кибернетикалық автоматтың мәтіндерді «адамша» аударуы бір адамның бойына біткен лингвистикалық білім дәрежесімен және басқа да табиғи мүмкіншіліктерімен шектелмеу керек, ол компьютер үшін жеткіліксіз. Болашақтың лингвистикалық компьютері бүкіл бір ұйымға тән, әлеуметтік кәсіби (профессионал) топтарға тән интеллектуалдық, мәдени-әлеуметтік және лингвистикалық тәжірибелерді өзіне игеруі қажет.

Сонымен, хабардың берілу жүйесінің салалық тілдермен жұмыс істеуі, табиғи тіліміздің коммуникативтік қызметін дамыта түсетіндігіне еш күмән жоқ деуге болады. Ал қазақ тіл білімінің компьютерлік саласының тиісті дәрежеде дамуын, оның қатаң ережеге бағынуынан, мәтін бойындағы тілдік бірліктерді тануға қажетті формалды қасиеттерінің айқындалуынан, омоним сөздер мен көпмағыналы сөздердің және синоним сөздер қатарын сирету мен реттеу мәселелерінің өз шешімін дұрыс табуынан байқауға болар еді.

Көптеген ғалымдардың пікірлерін қорыта келе, біз «Қазақ тілінің компьютерлік қоры» атты жүйенің түсінігі – көп тармақты әрі терең автоматтандырылған жүйені аңғартады демекпіз және оның атқаратын

міндеті ақпараттық барлау мен тіл білімі саласындағы (жалпы филология бағыттағы) зерттеушілік болуы тиіс. Ал ондай компьютерлік қор жасаудың методологиялық негіздері ретінде мынадай екі тезисті басшылыққа алғанды жөн санар едік. Оның біріншісі – тіл жөніндегі деректердің қандайы болмасын лексикографиялық пішінге келтірілу қажеттігі, ал екіншісі – лексикографиялық жағынан жүйеленген тілдік мәліметтердің барлығын да компьютерлік алгоритмдер тіліне лайықтаудың керектігі.

Тілдік бірліктердің (ақпараттардың) лексикографиялық пішінде көрініс табуы олардың сөздік түріндегі жұмыстардың идеясына жақындайды, ал мұндай тұжырым тілдік ақпараттарды шоғырландыратын бірегей түйін немесе түп қазық – жеке сөз, деп ұйғарумен бірдей. Шынында да, сөз өзінің ішкі-сыртқы қызметтері жағынан және ақпарат сақтағыштық қасиеттерінің негізінде белгілі бір жүйедегі сөз қорын танытатыны сөзсіз. Сондықтан сөздің белгілі мөлшерде өзіндік сегменттері болады және олар грамматикалық, стилистикалық, фонетикалық, контекстік т.б. пішінде көрініс табады, яғни ол барлық тілдік мәліметтер жүйесін сәулелендіреді деп ұйғарған жөн.

Егер біз ана тіліміздің компьютерлік қорын құру методологиясына орыс тілінің машиналық қорын жасау тәжірибесіне сүйенсек, онда мынадай қор құрамын сақтауымыз қажет:

1) қазақ тілі компьютерлік қорының негізгі сөзтізбегі (генеральный словник);

2) қазақ мәтіндерін суреттеу қоры (иллюстративный фонд);

3) қазақ тілінің терминологиялық қоры;

4) қазақ тілінің академиялық сөздіктері мен грамматика қоры;

5) қазақ тілінің лингво-статистикалық қоры;

6) қазақ тілінің лексикографиялық қоры;

7) қазақ тілі процессорларының қоры;

8) қазақ тіліне қатысты лингвистикалық алгоритмдер мен комплексті сипаттағы теориялық және тәжірибелік мәліметтерді бірлікте қамтитын бағдарламалар қоры [5, 9-10 бб.].

Аталған құрамдағы қазақ тілінің компьютерлік қоры тілдік мәліметтердің сала-салаға жүйеленген банк ретінде де және тұтыну қажеттігіне сай қайта жинақталған тұтас күйінде де пайдалану мүмкіндігі алдын ала ескерілуі керек.

Қазақ тілі жөніндегі деректерді жинаудың, сақтаудың, талдаудың және салыстыра зерттеудің жаңа әдістеріне көшу, сондай-ақ лингвистикалық дерек көздерінің автоматтандырылған сөздіктер мен грамматикалар тәрізді жаңа түрлерін игеру ісі өмірге жанасымды және тиімді екені сөзсіз. Әрине, бұл үшін біз жалпыфилологиялық дәстүрлер мен мәдениетке, тілді танып білуге, практикалық жаңа міндеттер тәжірибелерін мұқият ескеруге тиіспіз.

Біздің ұйғаруымызша, мұндай жүйелерде тұтас қалыптағы мәтіндер сақталуы қажет те және тұтынушының сұранысы бойынша сол мәтіндердің үзінділері, қор ішінде тұрған олардың нақты орны, сөзтұлғалардың

статистикасы, сөзнұсқағыштардың толық тізімдері, сөздерге телулі цитаталар, жиілікті және кері әліпбилі сөздіктер көрініс табуы қажет.

Компьютерлік қордың лексикалық сипаты түпнұсқа мәтіндерінің типіне байланысты деп санай отыра, қазақ тілінің жүйесін танытатын сөйленіс жағдаятына: көркем әдебиет тілін, газет-журнал тілін, ғылым мен техника тілін, ұйымдастыру-нұсқау құжаттары тілін, заң орындары мен сот мекемелер тілін, оқу орындары тілін, сауда, дәріхана, емханалардағы және т.б. сөйленіс түрлерін жатқызуды қолдар едік. Әрине, қордың бұл түріндегі компьютер жадына орналастыруға қиындық туғызатын сөйленіс түрлеріне жататындар: диспуттар, сот процестері, жиналыстар, күнделікті әңгімелесулер, телефон арқылы сөйлесу т.т. сияқты ауызекі сөйлесу тілінің үлгілерін жазып алу мәселесі.

Келесі бір қомақты мәтіндік бірлік – ол терминдер мәселесі. Қазіргі ғылыми-техникалық және іскағаздары мәтінінің негізгі өзегі терминдер екені айқын. Жалпы сөйлеу тілінің аясындағы сөздер емес, терминдердің тілдік қоршауы (мәнмәтіні) өзгеше бір тілдік жүйе, басқаша айтқанда, терминдер өрісі болып табылады. Осы аяда өмір сүретін термин үшін барлық ғылыми шығармалардың сөз қолданыс машықтары етене, ортақ болып келеді.

Осындай маңызды жұмыстардың келесі түрі – әр дәуірдегі мәтіндердің тілдік белгілерін ажыратып ала білу мәселесі. Осы мақсатта тіліміздегі диалектілік лексиканың біркелкі еместігін ескеретін болсақ, оларды компьютерлік қорға қосу жұмыстарының қазақ диалектологтары үшін ойланатын жақтары жоқ емес. Мәселен, олар бір жағынан ертедегі сөздер мен қолданыстардың ұлтқа, жазуға дейінгі үлгілері болып келсе, екінші жағынан басқа тілдерден енген сөздер, таралу өрісі шектеліп қалған сөздер немесе тіліміздегі жаңадан енген сөздер болуы да ықтимал.

Халық тілі сол халықтың жанды қазынасы болғандықтан, соның бүкіл болмысын барлық ерекшеліктерімен қоса қамту болашақ «Қазақ тілінің компьютерлік қорының» міндеті деп түсіну қажет сияқты.

Жоғарыда сөз болған қазақ тіліндегі диалектілер секілді, әрі синхрониялық, әрі диахрониялық сипатта күн кешетін тіл – фольклордың тілі және жер-су аттарының тілі немесе топонимика.

Осы аталған тіл үлгілерін жинастырып, тіліміздің компьютерлік қорын жасауда, оның бұл аталғандардан басқа да салалары мен арналары қамтылары сөзсіз.

Формалды пішіндегі қазақ тілінің компьютерлік қорын жасау мақсатымен байланысты күтілетін ең маңызды мәселелер мыналар деп білеміз:

а) тұтынушыға қажет болған белгілі бір сөзді қандай тіл көздерінен, мәтіндерден, тіл айырымдарынан іздестіру қажет;

ә) іздеуге тиісті сөз қарастыратын мәтінде қамтылған ба, жоқ па?

б) графикалық және фонетикалық пішіндерін қоса есептегенде сөздің мүмкін болған барлық түрлері мен тұлғасы айқындалып болды ма?

в) белгілі бір сөзге қатысты мәтін үзінділерінің (мәнмәтіндердің) жалпы аумағы, сипаты анықталды ма?

г) белгілі бір сөз: қарастырып отырған семантикалық және грамматикалық сипатта өткен дәуірлердегі, көне ғасырдағы мәтіндерде қолданылған ба, егер қолданылса, қандай сөздіктерден көрініс тапқан?

Міне, осы аталғандар қазақ тілінің компьютерлік қорын жасау мақсатындағы күн тәртібіне қойылатын ең маңызды мәселелер болуы қажет.

Жоғарыда сөз болған Совет түркологтары комитетінің XIV Пленумына қайта оралып, мәжілісте көтерілген негізгі мәселелерге тоқтала кетейік.

Алғашында мәжіліске қатысушылар, тілдің машиналық қорының анықтамасы төңірегінде пікір алмасып, соның нәтижесінде бір түйінге келді деуге болады.

ЭЕМ-ның (компьютердің) араласуымен зерттелетін жекеленген мәселелердің нәтижелерін, мысалы, жеке тұрып лексика-семантикалық, этимологиялық, морфологиялық, тіпті, синтаксистік есептерді шешетін машиналық бағдарламалардың жиынтығын немесе ЭЕМ-ның жадына енгізілген мәтіндер мен лексикографиялық материалдардың тұтастығын сол тілдің машиналық қоры, міне осы, деп есептеуге болмайтындығы баса көрсетілді.

Мәселен, ҚР БжҒМ А.Байтұрсынұлы атындағы Тіл білімі институтында компьютерлік лингвистика саласы бойынша біршама ғылыми жұмыстар орындалды. Оларды қысқаша былайша атап өтуге болар еді:

– көркем әдебиет, көсемсөз, ғылыми-көпшілік әдебиет пен оқулықтар мәтіндерінің, алғашында ЭЕМ-ның, кейіннен компьютердің жадына көптеп енгізіліп, олар бойынша арнайы компьютерлік бағдарламалар құрастырылуының негізінде әртүрлі жиілік сөздіктердің алынуы;

– М.Әуезовтің 20 томдық шығармалар жинағының мәтіндерінің компьютер жадына енгізілуі және соның негізінде жазушы тілінің жиілік сөздіктерінің жарық көруі;

– осы тәрізді жиілік сөздіктердің негізінде қазақ сөздерінің әр стильде және әр көлемді мәтіндерде қайталануының қосынды жиілігінің, сол мәтіндерді қамту статистикасының анықталуы;

– сөзтұлғалардың құрылымдық статистикасы мен мәтіндегі пайыздық салмақтарының әр сөз таптарына қатысты айқындалуы;

– қазақ тіліндегі негізгі сөз таптарының мәтін бойында кездесулерінің математикалық статистика мен ықтималдық теория заңдылықтарына бағыну-бағынбау жағдайларының анықталуы;

– қазақ тілінің «ТІЛ – ҚАЗЫНА» атты автоматтандырылған компьютерлік картотекалық қорының іске қосылуы;

– қазақ тіліндегі мәтін мазмұнын оның тұрпатына қарай ашудың (тұрпаттаудың, формалдаудың) негізгі принциптерінің айқындалуы;

– Тіл білімі институтында жоспарлы тақырыптар бойынша түзіліп жатқан түсіндірме, екі тілдік, бір тілдік сөздіктер мәтіндерінің компьютер жадында «сөздіктер қорын» құрап, тұрақты сақталуы.

Міне, осы жұмыстар түрлерінің жиынтығын біз қазақ тілінің толық мағынадағы компьютерлік қоры деп атай алмаймыз. Себебі оларды жеке мәселелерді шеше алатын ғана бағдарламалар мен мәтіндік файлдар

жиынтығы және олар бір-бірімен тығыз байланыста болмай, тұтынушы сұранысына бейімделмеген. Сонымен бірге мұндай жиынтық бір орталықтан басқаруға болатындай комплексті түрдегі автоматтандырылған күрделі жүйе емес.

Жоғарыда сөз болып жатқан Пленум мәжілісінде түркі тілдерінің көпаспектілі машиналық қорын жасау қолданбалы ілімдердің мұқтаждығына сай жүзеге асырылуы тиіс екендігі баса айтылған болатын. Өкінішке орай, мұндай жақсы бастама сөз жүзінде ғана қалып, түркі тілдерінің көпаспектілі машиналық қорын өмірге келтіру әлі де болашақтың ісіне айналып отыр.

Ал қазіргі жағдайда, ең алдымен, егеменді еліміздің мемлекеттік тіліне айналған ана тіліміз – қазақ тілінің компьютерлік қорын ғылыми тұрғыда құру – жаңаша ойлап, жаңаша шешудің ең тиімді жолы екндігі еш дау туғызбайды деп ойлаймын.

Енді түркі тілдерінің көпаспектілі машиналық қорын құруға қажетті деп есептелетін түрліше топтамаларына (блоктарына) жалпылама сипаттама бере отыра, оның төменде аталатын тарамдарын ұсынып отырмыз.

1. Алдымен құрастыруды қажет ететін сала – лексикалық блок. Ол орфографиялық, орфоэпиялық, түсіндірме, терминологиялық, жиілік және аударма сөздіктерді қамтитын көпшілік қолды лексикографиялық жүйе болып табылады.

2. Грамматикалық блок – лексикалық және грамматикалық морфемалардың тізбелерінің жиынтығынан, сондай-ақ синтаксистік модельдерді (сызбалық) қамтитын тілдік материалдардан құралады.

3. Фонетикалық блок – тілдің фонемалық құрамын анықтайды, әр фонеманың айырым (дифференциал) белгілерінің анықтамалары беріледі.

4. Морфонологиялық блок:

а) тілдік бірліктердің фонемдік құрылысының ерекшеліктері жөніндегі деректер қамтылады. Мәселен, негіздер, форманттар, сөздердің жекелеген кластары (бейнелеуіш сөздер) т.б.;

ә) тілдің сингармонизм заңдылығы, сондай-ақ дауыссыз дыбыстардың ассимиляциясы, қатаңдауы мен ұяңдауы жайындағы мәліметтер ескеріледі;

б) екпіннің орны көрсетілген түбір негіздер мен сөзформалар жинақталады [5, 14-б.].

Міне, осы аталған салалардың бәрін жан-жақты қамти отыра, қажетті деген сұранысқа толық та және тұжырымды жауап алу үшін біз соңғы ғылыми-технологиялық жетістіктерді пайдалана отырып, әрі жүйелі, әрі сапалы түрде көрініс табатын ғылыми тұрғыдағы қазақ тілінің компьютерлік қорын іске қосуымыз қажет.

Мынадай сұрақ туындауы мүмкін: «Ашық жүйе ретіндегі қазақ тілінің компьютерлік базасы» құрастырылып болғаннан кейін ол тек архив тәрізді қызмет ете ме, әлде ол тіл жөніндегі жаңа білімдер қосатын, жаңа мәліметтер тауып беретін ғылыми зертхана тәрізді бомақ па?

Бұл ретте айтарымыз, екі жақты құбылыстың соңғысы алғашқы қисынның барлық қасиеттерін иемденетіндігі себепті, оның берер мүмкіншіліктері де молырақ болмақ. Демек, «Ашық жүйе ретіндегі қазақ

тілінің компьютерлік базасы» бізге белгілі тілдік заңдылықтар жөніндегі мәліметтерді бойына жинақтаған тілдің синхронды және диахронды процестерін модельдей алатын әрекет үстінде болуы қажет. Сонда ғана компьютерлік тілдік қор түрлі деңгейдегі тіл қабаттарымен жүйелі негізде қызмет ететін болады. Басқаша айтқанда, «Қазақ тілінің компьютерлік базасы» дегеніміз қазақ тілі жайлы барлық мәліметтерді жинақтаушы банк, бірақ ол «ақылды», мамандандырылған білім қоры. Оны өзімізге дәстүрлі кітап ретінде де пайдалануға болады және зерттеуші өз ұсыныстарын белгілі бір мақсаттарда жүзеге асыра алады.

Осы саладағы белгілі ғалымдардың тұжырымдауынша, компьютерлік тілдік қор дегеніміз ғылым адамының өз зерттеу нысанына жаңаша тұрғыда көз салудың мүмкіндігі болып табылады. Мұндай тілдік қор неғұрлым қомақты болса, солғұрлым тіл құрылысының сырын тереңірек аша алады, сөйтіп зерттелетін нысан жөніндегі түсініктердің шеңбері кеңиді, адамның білім өрісіндегі «ақтаңдақтардың» бедер-бейнесі айқындала түседі.

Біздер де осы айтылғандарға қосыла келе, өз атымыздан «Қазақ тілінің компьютерлік базасы: лексикалық қордың ашық жүйесі» зерттеуші адамның қалып-қабілетін әлденеше есе арттырады, шығармашылық қуат көздері ашыла түседі, сөйтіп бұл жаңа мүмкіндіктер қазақ тілінің жүйелілік қасиеттерін жетілдіруге және тіл жүйесін мұқият тануға жұмсалар еді деп тұжырымдар едік. Әрине, бұл аталып отырған тілдік қорды толығымен өмірге келтіру үшін оның жоба-жолдары мемлекеттік дәрежеде, яғни ҚР Білім және ғылым министрлігі, Тіл комитеті деңгейінде қаралуы қажет. Себебі, мұндай жалпы филологиялық, ұлттық және мәдениет дәрежесіндегі күрделі жүйені іске қосу – қажетті мамандарды (жоғары дәрежелі инженер-бағдарламашыларды, компьютерлік лингвистика саласының мамандарын және т.б.) және тиісті қаражат мүмкіндігін шешуді талап етеді.

Сөз соңын, машиналық қор жайындағы идеяны осыдан ширек ғасыр бұрын ең алғаш рет ұсынған информатика саласының белгілі ғалымы А.П.Ершовтың сол кездердегі айтқан сөзімен қорытындылайық.

А.П.Ершов 1978 жылы жарық көрген «Іс қағаздары прозасының феномены: диалогтар жүйесін жасау методологиясына» атты мақаласында: «Егер орыс тілінің машиналық қорын жасау мақсаты ойдағыдай орындалса, модельдеу мен алгоритмдерді құрастыру саласындғы қол жеткен табыстарымыздың бәрі академиялық жаттығулар қалпында қалып қоймай, іс жүзіне асқан болар еді, мұндай өте-мөте қажетті проблеманың шешілуі ғылым, мәдениет және қолданбалы техникалық тәсілдер тұрғысынан мейлінше құнды болмақ. …Тілдің машиналық қорын жасау ісі өндірістік лингвистикалық жүйелерді іске қосудан озық болуы тиіс, өйткені тек осы жағдайда ғана толып жатқан қажетсіз қайталаулардан құтылып, тілдің жанды өзегін озбырлар мен дүмшелердің қиянатынан аман сақтап қалуға болады», - деп тұжырымдайды.

Әдебиет

1. *Гузев В.Г., Пиотровский Р.Г., Щербак А.М.* О создании машинного фонда тюркских языков // Советская тюркология. 1988. №2. С. 98-101.

2. *Ершов А.П.* К методологии построения диалоговых систем: феномен деловой прозы. Препр. №156, ВЦ СО АН СССР. Новосибирск, 1979.

3. *Ершов А.П.* Машинный фонд русского языка: внешняя постановка // Машинный фонд русского языка: идеи и суждения. М.: Наука, 1986. 240 с.

4. *Андрющенко В.М.* Концепция и архитектура машинного фонда русского языка. – М.: Наука, 1989. –200 с.

5. *Жұбанов А.Қ.* Қазақ сөзінің «**Тіл – қазына**» деректер базасы және оның теориялық негіздері. База данных «**Тіл – қазына**» казахского слова и ее теоретические основы. – Алматы: «Арыс» баспасы, 2009. – 304 бет.

ҚАЗАҚ ЛЕКСИКОГРАФИЯСЫНЫҢ «ТІЛ – ҚАЗЫНА» АТТЫ ЭЛЕКТРОНДЫҚ БАЗАСЫ ЖАЙЫНДА

ҚР БҒМ ҒК А.Байтұрсынұлы атындағы Тіл білімі институты. Көрнекті түркітанушы ғалым, филология ғылымдарының докторы, профессор Бабаш Әбілқасымовтың 80 жылдық мерейтойына арналған «Қазақ тілінің тарихы мен түркітану мәселелері» атты халықаралық ғылыми-теориялық конференция материалдары. Алматы-2011, 225-228-бб.

Тілдік деректердің лексикографиялық пішінде көрініс табуы оларды сөздік түріндегі жұмыстардың идеясына жақындатады, ал мұндай тұжырым тілдік ақпараттарды шоғырландыратын бірегей түйін немесе түпқазық – **жеке сөз** деп ұйғарумен бірдей. Шынында да, сөз өзінің ішкі-сыртқы қызметтері жағынан және ақпарат сақтағыштық қасиеттерінің негізінде белгілі бір жүйедегі сөз қорын танытатыны сөзсіз. Сондықтан сөздің белгілі мөлшерде өзіндік сегменттері орын алады және олар грамматикалық, стилистикалық, фонетикалық, контекстік т.б. пішінде көрініс табады, яғни ол барлық тілдік мәліметтер жүйесін сәулелендіреді деп ұйғару қажет [1, 4-б.].

Тілдік деректерді бір-бірімен тығыз байланыстағы, тұтынушы сұранысына бейімделген және бір орталықтан басқарылатын кешенді түрдегі автоматтандырылған күрделі жүйе құру үшін, біз соңғы ғылыми-технологиялық жетістіктерді пайдалана отырып, әрі жүйелі, әрі сапалы түрде көрініс табатын ғылыми тұрғыдағы қазақ тілі деректерінің компьютерлік базасын құрастыруды мақсат еттік.

Міне, сол себепті қазақ тілінің лексикалық қорын ашық жүйе ретінде қарастыра отырып, қазақ сөзінің **«Тіл – қазына»** атты деректер базасын құру мәселесі ҚР БжҒМ А.Байтұрсынұлы атындағы Тіл білімі институтының кезекті зерттеулерінің күн тәртібіне енді. Бұл базаға алғашқы енгізілетін нысан ретінде көп жылдық тарихы бар және аса ұқыптылықпен жинақталған көлемі, шамамен алғанда, 5 миллиондай болатын қазақ тілінің картотекалық қоры алынды [1, 5-б.].

Қазақ сөзінің **«Тіл – қазына»** атты деректер базасы үлкен көлемді картотека мәтіндерімен бірге көп көлемді сөздіктерге де қатысты ақпараттық-сұраныс жүйесі түрінде көрініс табуы мақсат етілді. Ондай деректер базасының нысаны ретінде келесі сөздіктердің мәтіндері алынды: 1) «Абай тілі сөздігі» (6 мың сөз, 1 мыңдай сөз тіркес), 2) бір томдық «Қазақ тілінің сөздігі» (50 мың сөз) және 3) жаңадан құрастырылып жатқан 15 томдық «Қазақ әдеби тілінің сөздігі» (120 мың сөз).

Келесі бір қомақты мәтіндік бірлік – ол терминдер мәселесі. Қазіргі ғылыми-техникалық және іскағаздары мәтінінің негізгі өзегі терминдер екені айқын. Жалпы сөйлеу тілінің аясындағы сөздер емес, терминдердің тілдік қоршауы (мәнмәтіні) өзгеше бір тілдік жүйе, басқаша айтқанда, терминдер өрісі болып табылады [1, 11-б.].

Сондықтан қазақ тілінің терминологиялық қорына салалық ақпараттық жүйелер мен терминологиялық стандарттарда тіркелген барлық лексикалық бірліктерді қамту да негізгі мақсаттардың біріне айналды. Дәлірек айтсақ, ҚР Үкімет жанындағы Мемлекеттік терминология комиссиясы бекіткен және республикалық «Рауан» баспасынан шыққан ғылым мен мен мәдениет салаларының көпшілігін қамтитын қазақша-орысша және орысша-қазақша терминологиялық сөздіктердің 31 томдық сериалы қазақ сөзінің «**Тіл – қазына**» базасына енгізілді [1, 36-37-бб.].

Сол сияқты қазақ сөзінің «**Тіл – қазына**» атты электрондық базасына, ақпарат және қоғамдық келісім министрлігі ұсынып, «Мектеп» баспасынан жарық көрген «Қазақ тілі терминдерінің салалық ғылыми түсіндірме сөздіктерінің» 10 томдық топтамасының мәтіндері де енгізілді [1, 38-б.].

Осы аталған тіл үлгілерін жинастырып, тіліміздің компьютерлік қорын жасауда, оның бұл аталғандардан басқа да салалары мен арналары болашақта қамтылары сөзсіз.

Ал осындай деректер бойынша компьютерлік база құру мақсатында мына міндеттер жүзеге асырылды:

1) картотека мен сөздіктер мәтінін электрондық пішінге келтіру;

2) компьютерлік база құру жолындағы, объективті-реляционды деректер базасын, СУБД мен Web құралдарын қамтитын әлемдік тәжірибе негізінде компьютерлік алгоритмін құру және оның жасанды тілдегі бағдарламасын жазу;

3) қазақ тілінің картотекалық және сөздіктерінің **«Тіл – қазына»** атты деректер базасын әрі бөлек пішінде, әрі олардың объективті-реляциялық база арқылы бір-бірімен тығыз байланыста болуын ескеру;

Бұл аталған жұмыстардың даму барысы нәтижесінен қазақ тілінің мәтін семантикасы мен лексикасы аясындағы ғылыми зерттеулердің және түсіндірме сөздіктердің маңыздылығы арта түседі және мұның ұлттық «жасанды интеллектіні» құрастыру ісіне тікелей қатынасы бар. Яғни «қазақша ойлай алатын» зияткерлі роботтарды басқаруда да аса маңызды демекпіз.

Картотекалық қорды автоматтандыру жүйелерін, автоматты сөздіктер мен деректердің терминологиялық банкін лексикографиялық жұмыстарды

автоматтандыру жүйесі немесе автоматтанған лексикографиялық жүйе деп қарастыруға әбден болады.

Автоматтанған лексикографиялық жүйе төмендегідей лексикографиялық амалдарды орындау қажет:

- сөзтізбені іріктеу және бірнеше екітілді сөздіктер үшін біркелкі сөзтізбе қалыптастыру;
- сөзтізбені автоматты түрде морфологиялық ақпаратпен қамтамасыз ету;
- иллюстративтік материалды таңдау;
- алдын ала белгіленген мәтіндерді топтастыруға қатысты стилистикалық таңбалардың сәйкестігін бақылау және т.б.

Автоматты сөздіктерді сөздіктердің жеке бір түрі деп ұғынудан гөрі, белгілі мақсат үшін пайдалану сөздігі деп түсінген жөн сияқты. Сондықтан, тұтынушы-лексикограф автоматты сөздікте орын алатын мақала ақпаратын әртүрлі көзқараста ұғынуына мүмкіндігі бар, мысалы, егер деректер базасындағы тиісті грамматикалық (фонетикалық) ақпаратпен жабдықталған сөздік мақаланың тақырыбына ғана назар аударатын болсақ, ол сөздікті орфографиялық немесе орфоэпиялық, фонетикалық немесе грамматикалық және т.б. сөздік деп ұғынуына болады.

Ұсыныс ретінде айтқымыз келгені, гуманитарлық зерттеулер үшін арнайы компьютерлік орталық ашылғаны жөн болар еді. Сонымен бірге талай мәрте қолданыс табатын мәтіндер, сөздіктер, ақпараттық материалдар, қоғамдық пікірге қатысты сұрақ-жауаптар және т.б. материалдар да әрдайым қорланып тұруы керек. Қазақ тілінің сөздіктерін, анықтағыштарын, энциклопедияларын, жинақтарын, оқулықтар мен оқу құралдарын, белгілі оқырман аясына бейімделген әдеби және ғылыми-көпшілік шығармаларын компьютер жадында сақтау технологиясын меңгерудің маңыздылығын айрықша атауға болады.

Әрине, қазақтың гуманитарлық зерттеулеріне арналған компьютерлік орталық гуманитарлық білім иелерін, яғни маман емес тұтынушыларды дайындайтын оқыту орталығына айналғаны да жөн болар еді.

Біз жаппай автоматтандыру заманында өмір сүретіндігімізден, қоғамның әрбір мүшесі, белгілі дәрежеде, «компьютерді пайдаланушы» атына ие болатыны белгілі. Сондықтан білімнің гуманитарлық саласына, әсіресе қазақ тіл біліміне, компьютер мүмкіндігін толығымен пайдалану егемен еліміздің болашақ дамуына қосқан зор үлес болар еді [2; 3].

Қазақ тінінің **«Тіл-қазына»** ұлттық компьютерлік базасын толыққанды құрастыру дегеніміз, ол тек тілдің қолданбалы жақтарын қамту ғана емес, сонымен бірге теориялық зерттеулер мен тәжірибелік жұмыстар үшін де аса құнды және қазақ тілінің анықтамалық қызметі тиімділігін арттырады.

Қазақ тілінің «Тіл-қазына» ұлттық компьютерлік қоры лексикография саласының дамуына, жазушылар тілін, дара тілдік шығармашылықты зерттеуге қозғаушы күш болары сөзсіз. Сонымен бірге, бұқаралық сөйлеу тілін, баспа тілін, қоғамның соңғы талаптарын реттеуге себепші болатын радио мен теледидар тілдеріне жүйелі түрде зерттеу жүргізуге мүмкіндік

туады [2; 3].

Зерттеулдің формалды және математикалық әдістерін, олардың нәтижелер берілісін барынша толық меңгерген жөн. Себебі, тек формалды көрініс тапқан және математикалық дәлдікте берілген нәтижелер компьютерлік «Тіл-қазына» қорының игілігі бола алады және осы қор арқылы басқа зерттеушілер мен тұтынушылардың да түсінуіне қол жетеді. Ең алдымен бұл сөздіктер базасы деректеріндегі әр түрлі сөздіктерді тікелей салыстыру мен біріктіру негізінде ғана мүмкін болатын, лексикографиялық деректердің формалды көрінісіне (берілісіне) қатысты келеді.

Бір тілден екінші тілге автоматты аударма жасау жүйелеріндегі формалды әдістер қазақ тілі мен шет тілдері арасындағы өзара әрекеттестікте аса қажет болуы мүмкін. Мәселен, қазақша-орысша, орысша-қазақша автоматтанған сөздіктер құрастыруда, автоматты аударма жасауда және т.б. екі тіл аралық зерттеулерде аса пайдалы. Ол үшін автоматтанған аударма мен автоматты лексикография салаларындағы қол жеткен нәтижелерді қазақ тіл лексикография тәжірибесіне жедел түрде енгізу қажет болады.

Ұсыны ретінде тағы да бір айтарымыз: қазақ тілінің «Тіл-қазына» ұлттық компьютерлік қорын құрастыру мен әрі қарай дамыту ісіне жас маман-зерттеуші буынының ат салысқаны жөн болар еді. Ол үшін қазақстандағы, жоғарғы оқу орындарында филология факультеттерінде құрылымдық, қолданбалы және математикалық лингвистика бойынша мамандақтар мен кафедралар ашылғанын қолдар едік.

Қазақ сөзінің **«Тіл – қазына»** атты деректер базасын толық сипатта өмірге келтіру үшін оның жоба-жолдары мемлекеттік дәрежеде, яғни ҚР Білім және ғылым министрлігі, Тіл комитеті деңгейінде қаралуы қажет сияқты. Себебі, мұндай жалпы филологиялық, ұлттық және мәдениет дәрежесіндегі күрделі жүйені іске қосу – қажетті мамандарды және тиісті қаражат мүмкіндігін шешуді талап етеді [2; 3].

Сөз соңында, машиналық қор жайындағы идеяны осыдан ширек ғасыр бұрын ең алғаш рет ұсынған информатика саласының белгілі ғалымы А.П.Ершовтың сөзімен қорытындыласақ, ол былайша айтқан болатын: «Егер орыс тілінің машиналық қорын жасау мақсаты ойдағыдай орындалса, модельдеу мен алгоритмдерді құрастыру саласындғы қол жеткен табыстарымыздың бәрі академиялық жаттығулар қалпында қалып қоймай, іс жүзіне асқан болар еді, мұндай өте-мөте қажетті проблеманың шешілуі ғылым, мәдениет және қолданбалы техникалық тәсілдер тұрғысынан мейлінше құнды болмақ. …Тілдің машиналық қорын жасау ісі өндірістік лингвистикалық жүйелерді іске қосудан озық болуы тиіс, өйткені тек осы жағдайда ғана толып жатқан қажетсіз қайталаулардан құтылып, тілдің жанды өзегін озбырлар мен дүмшелердің қиянатынан аман сақтап қалуға болады» [4].

Әдебиет

1. *Жұбанов А.Қ.* Қазақ сөзінің **«Тіл – қазына»** деректер базасы және оның теориялық негіздері. База данных **«Тіл – қазына»** казахского слова и ее теоретические основы. – Алматы: «Арыс» баспасы, 2009. – 304 бет.

2. *Жұбанов А.Қ.* Компьютерлік лингвистикаға кіріспе: Оқу құралы / Қазақ университеті. –Алматы, 2007. – 204 б.

3. *Жұбанов А.Қ.* Тіліміздің компьютерлік қорын жасау керек //Ана тілі. – 2003. – 13 қараша (№ 45).

4. *Ершов А.П.* К методологии построения диалоговых систем: феномен деловой прозы. Препр. №156, ВЦ СО АН СССР. Новосибирск, 1979.

ҚАЗАҚ ТІЛІ КОРПУСЫНДАҒЫ ЗАТ ЕСІМ СӨЗДЕРГЕ ЛЕКСИКА-МОРФОЛОГИЯЛЫҚ ДЕҢГЕЙДЕГІ БЕЛГІЛЕНІМ ЕНГІЗУ БАҒДАРЛАМАСЫ

ҚР БҒМ ҒК А.Байтұрсынұлы атындағы Тіл білімі институты. Көрнекті түркітанушы ғалым, филология ғылымдарының докторы, профессор Бабаш Әбілқасымовтың 80 жылдық мерейтойына арналған «Қазақ тілінің тарихы мен түркітану мәселелері» атты халықаралық ғылыми-теориялық конференция материалдары. Алматы-2011, 228-232-бб.

Тілдік нысандарға қолданылатын математикалық әдісті интуитивті түрде тұжырымдалған, сол сияқты толық шешімі жоқ мәселелердің логикалық тұжырымдалуына бағынатын және алгоритмдік шешімі бар бір немесе бірнеше қарапайым математикалық есептермен ауыстыру деп ұғыну қажет. Тілдік мәселені осылайша математикалық жолмен шешу – лингвистикалық нысанның математикалық моделіне (үлгісіне) көшуді қажет етеді [1, 6-б.]. Тілдің аса күрделі түрдегі сан алуан параметрлі көп қабатты жүйе екенін ескерсек, онда ықтималдықтар теориясы, математикалық статистика, ақпарат теориясы және т.б. ғылыми салалар әдістерін қолданудың мүмкіндігі мол.

Қазақ тілінің қолданбалы саласының, дәлірек айтқанда, статистикалық лингвистиканың алдына қойған мақсаттарының бірі – сөз табының лексикалық (семантикалық), грамматикалық (морфологиялық және синтаксистік) белгілері арқылы, яғни осы үш белгінің негізінде статистикалық зерттеулер жүргізіп, сөздің мағынасын, морфологиялық формасын (тұлғасын) және синтаксистік қызметін сандық деректер арқылы тани білу. Соның нәтижесінде мәтін мазмұнын ашудың формальды жақтарын айқындау.

Сөз таптарының ішкі жіктелген бөліктерінің түрлері тілде біреуі аз не көп қолдануы мүмкін. Статистикалық зерттеу тәжірибесі бойынша олардың қолдану жиілігін анықтау үшін кейбір шараларды алдын ала белгілейтін бағдарламаларды іске қосу қажет.

Мысалы, әр түрлі тілдік стильдердегі сөз таптарының қатынастық сипатының статистикасын білу керек болса, әр сөз табына қатысты шартты белгіні (белгі-кодты) қоюдың зерттеушіге ыңғайлы жолдарын ойластыруымыз керек. Себебі, сөз таптарын не басқа тілдік бірліктерді мәтін ішінде түр-тұрпатына қарай формальды әдіспен ажырату мүмкіндігі әлі де болса зерттеу аясынан тыс қалып жүр. Сондықтан статистикалық санақтарды компьютер арқылы жүргізуде аталған тілдік бірліктерді және олардың ішкі бөліктерін шартты белгілерге сәйкестендіру (кодтау) танымдық рөл үшін аса қажетті. «Кодтау» принципі алдын ала ойластырылған әмбебаптық сипатта болғаны жөн.

Шартты белгі-кодтарды қоюдың мүмкіндігі мол, дегенмен оларды әркім өз қалауынша белгілегеннен гөрі біріздiлікке (стандартқа) ұмтылған дұрыс деп саналады. Әр сөз табының іштей морфологиялық, синтаксистік және семантикалық тармақтану түрлері мен жасалу жолдары да әр зерттеушінің алға қойған мақсатына сай әр түрлі сипатта болуы мүмкін. Сондықтан әрбір сөзге, сөзтұлғаға, сөзқолданысқа, сөзтіркеске немесе мәтіннің одан да үлкен бірліктеріне қойылатын «белгі-код» әр түрлі грамматикалық ақпаратты өзінде сақтай алады. Шартты белгі-кодтарды қабылдау тек бір ғана тілдік бірліктің бірнеше ақпараттарын сәйкестендіретін «ұялар» тізбегі ретінде ұйымдастыруға болады.

Енді осындай белгілеу әдісін күрделендіріп, әр сөз табының морфологиялық типтері мен құрылымы жағын ажырату, жасалу жолдарын ескеру, яғни сол тілдік бірлікке қатысты басқа да грамматикалық ақпараттарды барынша қамту және олардың әрбіреуінің статистикасын анықтау мақсатымен, әрі қарайғы баяндауымызда «кодтаудың» кейбір үлгілеріне толығырақ тоқталамыз.

Орыс тілінің морфологиялық сипатына тән сөз таптарын шартты түрде белгілеу (кодтау) ресейлік ғалым Б.Н.Головиннің «Язык и статистика» атты кітабында да сөз болады [2].

Ал морфологиялық ақпарат белгілерін сәйкестендіру бағдарламасын құрудағы біз ұсынып отырған үлгінің ерекшелігі – қазақ тілі сөз таптарының ішкі бөліктерінің құрылымдық сипаты мен жасалу жолдары ескеріліп, әмбебаптық сипатты көздейді [3].

Бағдарлама жасау ісінің нысаны ретінде – зат есім, етістік, сын есім, сан есім, есімдік сияқты қазақ тілінің негізгі сөз таптары алынды. Аталған сөз таптарына морфологиялық ақпаратты белгі-код үлгісін сәйкестендіру бағдарламасы тиісті кестелер арқылы көрініс тапты. Ең алдымен қазақ мәтіндеріндегі зат есім сөздерге белгі-код сәйкестендіру бағдарламасын жеке қарастырайық.

Зат есім сөздерге ғана тән морфологиялық ерекшеліктер жоқ емес. Олар өздерінің лексика-семантикалық сипаттарына қарай арнайы жалғаулар арқылы түрленіп және жұрнақтар арқылы тұлғалық өзгеріске ұшырап, сөйлемдегі өзге сөздермен еркін қарым-қатынасқа түсіп отыратыны белгілі. Зат есім сөздердің морфологиялық, синтаксистік және семантикалық белгілерін ұштастыра қарағанда ғана олардың сыр-сипатын толық ашуға

болады. Сол белгілер қатарына олардың мәтін ішінде жиі не сирек қолдану сипаты да жатады.

Қазіргі қазақ тілінің грамматикасында зат есім сөздерге ғана тән семантикалық және грамматикалық ерекшеліктері бар мынадай топтар бөліп қарастырылады: адамзат (кімдік) және ғаламзат (нелік) есімдер, жалқы есімдер, көптік мәнді есімдер, эмоциялы-экспрессивтік зат есімдер, көмекші есімдер. Зат есім сөздерді олардың морфологиялық сипаты, яғни құрылымдық, жасалу тәсілі және түрлену жүйесі арқылы да топтап қарастыруға болатыны да белгілі. Міне, осылайша зат есім сөздерді топ-топқа бөліп қарастыру қазақ тілін компьютерлік және статистикалық лингвистика әдістерімен зерттеуде аса маңызды деуге болады.

1-кесте **Зат есім сөздерге лексика-морфологиялық ақпаратты сәйкестендіру бағдарламасы**

Белгі-код орны	***ЗАТ ЕСІМ СӨЗДЕРДІҢ ЛЕКСИКА-МОРФОЛОГИЯЛЫҚ СИПАТЫ***	Белгі-код
1	Зат есім сөз (сөзтұлға)	ЗТ/
2	Адамзат есімдері *(кім?)*	020
	Ғаламзат есімдері *(не?)*	021
3	Жалқы есім: а) кісі аттары (ономастика); ә) географиялық атаулар (топонимика)	030
	Жалпы есім	031
	Аралас мағыналы *(ай, күн, жер)*	032
4	Көптік мәнді емес (жалғаусыз қарастырғанда)	040
	Табиғи қос мәнді *(аяқ, құлақ, етік)*	041
	Көптік мәнді (жалғаусыз)	042
5	Реңсіз зат есім	050
	Эмоциялы-экспрессивтік реңді зат есім *(Әкей, Сәулеш)*	051
6	Көмекші зат есімге жатпайтын зат есім	060
	Көмекші зат есім *(алды, арты, қасы)*	061
7	Түбір морфема зат есім	070
	Туынды зат есім (түбір+жұрнақ)	071
	Күрделі зат есім: а) біріккен сөз	072
	ә) қосарланған сөз	073
	б) құрама сөз	074
	в) қысқарған сөз	075
8	Зат есімнің түрленуі:	
	Зат есімнің жалғаусыз түрі *(кім? не?)*	080
	Көптік жалғаулармен*: -лар, -лер,-дар,-дер,-тар,-тер*	081
	Тәуелдік жалғаулар арқылы: *-нікі,-дікі,-тікі -м,-ым, -ім; -ң, -ың, -ің, -ңыз, -ңіз, -ыңыз, - іңіз; -сы, -сі, -ы,-і*	082
	Септік жалғаулар арқылы: ілік *(кімнің? ненің?);* барыс *(кімге? неге? қайда?);* табыс *(кімді? нені?);* жатыс *(кімде? неде?);* шығыс *(кімнен? неден?);* көмектес *(кіммен? немен?).*	083
	Жіктік жалғау арқылы: жекеше*: -мын, -мін, -сың, -сің, -сыз, -сіз;* көпше*: -мыз, -міз.*	084

Қазақ тіліндегі мәтіндерде ең жиі қолданыс табатын сөз таптары *зат есім* мен *етістік*. Олай болатын болса, олардың мәтін ішінде морфологиялық, синтаксистік жолмен түрленіп қолдануының да статистикасын танып білу «Қолданбалы лингвистика» пәнінің мақсаттарының бірі. Біздің баяндауымыздағы негізгі ұстаным – мәтін ішінде кездесетін *сөз*, *сөзтұлға* не *тұрақты сөз тіркеске*, яғни мәтін бірлігі деп аталатын – *сөзқолданысқа* сәйкес келетін грамматикалық ақпараттар *орны («ұясы»)* мен сол ақпаратты танытатын шартты *белгі-кодты* тұрақтандыру. Қандай да болмасын тілдік бірліктің, мысалы, ең алдымен қарастырайын деп отырған – *зат есім* сөздердің әрбіреуіне сәйкестендірілетін шартты параметрлерді («ұя» орны мен белгі-код) алдын ала жан-жақты ойластырып алғанды жөн санаймыз (***1-кестені*** қара).

Айталық, *зат есім* сөздерге сәйкестендіретін (тиісті) «ұя» саны немесе топтама саны – сегізге тең делік, яғни n=1, 2, 3, 4, 5, 6, 7, 8. Ал әрбір «ұяда» жазылатын шартты *белгі-код* саны (топтама саны) мен «код мәні» *зат есім* сөздердің сол «ұяда» қарастырылатын сипатына қарай өзгеріп отыруы мүмкін.

Егер осы айтылғандарды қамтитын мәліметтер кестесін *«бағдарлама»* деп атасақ, *зат есім* сөз табына қатысты қысқаша топтама *1-кестедегі «бағдарлама»* түрінде көрініс табады:

Топтамадағы мәліметтердің кейбіреулеріне түсініктеме берейік.

Аталған кесте үш «бағанадан» тұрады:

1) *шартты белгілер орны* (топтама саны);

2) *зат есім сөздердің лексика-морфологиялық сипаты*;

3) *шартты белгі-код.*

Әрбір сөз табына қатысты «ұялар» топтамасының ең алғашқысы сол сөз табының қысқарған атымен белгіленеді (кодталады) және *1-ші орынға* ие болады. Сонымен, «ұялар» тізбегінің ең алғашқысы (бірінші бағана) «1» санына тең (бірінші топтама), осы жатық жолдағы екінші бағана сөз табының сипаты жайлы мәлімет, яғни – *зат есім* атауы және үшінші бағана – мәліметтің «белгі-код» мәні – «ЗТ». Сол сияқты шартты белгілер орнының «2» санында, яғни екінші топтамасында *зат есім* сөздердің екі түрлі сипаттамасы берілген. Олардың біріншісі – *адамзат есімдері (кім?)* және екіншісі – *ғаламзат есімдері (не?)*. Шартты белгі орнында, яғни 2-ші орында қарастырылатын кезекті *зат есім* сөздің қай сұраққа *(кім? не?)* жауап беруіне қарай үшінші бағанадағы оған сәйкес келетін белгі-кодтың біреуі ғана таңдалып алынады. *1-кестеде* зат есім сөздерге қатысты осы тәріздес 8 топтама қарастырылды.

Сол топтамалардың енді 7-ші орынға қатыстысын сөз етсек, жетінші топтамада *зат есім* сөздің алты түрлі сипаты аталған: *1. Түбір морфема зат есім; 2. Туынды зат есім (түбір+жұрнақ); 3. Күрделі зат есім: а) біріккен сөз; ә) қосарланған сөз; б) құрама сөз; в) қысқарған сөз.*

Үшінші бағанада олардың белгі-код мәндері үш орынды сандар: 070, 071, 072, 073, 074, 075 сандары арқылы берілген (*1-кестені* қараңыз).

Жоғарыда 2-ші орынға байланысты айтылған сияқты, бұл жолы да зерттеуге алынған кезекті сөзді (сөзтұлғаны) шартты түрде «кодтау» кезінде

7-ші орынға сәйкес келетін (070 пен 075 аралығы) алты белгі-код мәндерінің ішінен сөзтұлға сипатына тән келетін біреуі ғана алынады. Кестеде келтірілген басқа топтама орындарына да осылайша түсініктеме беруге болады.

Енді мысал ретінде кейбір зат есім сөздерді *1-кесте* деректері бойынша шартты белгі-код түрінде жазайық. Мысалы, келесі сөйлем ішіндегі зат есім сөздерді бөліп алып, белгі-код арқылы жазу керек: *Асан мектепке барарда кітап, қалам, дәптер және басқа да керек-жарақтарын аса ұқыптылықпен сөмкесіне салды.* Бұл сөйлемдегі зат есім сөзтұлғалар мыналар: *Асан; мектепке; кітап; қалам; дәптер; керек-жарақтарын; сөмкесіне.*

Енді, шартты белгі-кодтарды осы сөздерге сәйкестендірейік. Яғни әрбір аталған *зат есім* сөздердің сегіз орындық «ұяларына» тиісті *белгі-код* мәнін қойып шығу керек деген сөз. Олардың барлығының да бірінші орындарында «ЗТ/» белгі-код тұратыны айғақ.

Енді *Асан* сөзін бөлек алып қарастырсақ, 2-ші орында – «020» (адам аты болғандықтан), үшіншіде – «030», төртіншіде – «040», бесіншіде – «050», алтыншыда – «060», жетіншіде – «070», сегізіншіде – «080». Сонымен, сөйлемдегі *Асан* сөзі мен басқа да зат есім сөздердің «кодтық» жазылуы төмендегідей болады:

Асан:	ЗТ/	020	030	040	050	060	070	080
мектепке:	ЗТ/	021	031	040	050	060	070	083
кітап:	ЗТ/	021	031	040	050	060	070	080
қалам:	ЗТ/	021	031	040	050	060	070	080
дәптер:	ЗТ/	021	031	040	050	060	070	080
керек-жарақтарын:	ЗТ/	021	031	042	050	060	073	084
сөмкесіне:	ЗТ/	021	031	040	050	060	070	083

Көрсетілген жолмен мына зат есімдерге қатысты сөзтұлғаларды белгі-кодпен сәйкестендірейік: *ғалым, Асандар, құлағым, Алматыға, жазушы, саңырау құлақтан, жаным-ай, жүрісіме.*

ғалым	ЗТ/	020	032	040	050	060	070	080
Асандар	ЗТ/	020	030	040	050	060	070	081
құлағым	ЗТ/	021	032	041	050	060	070	082
Алматыға	ЗТ/	021	030	040	050	060	070	083
жазушы	ЗТ/	020	032	040	050	060	071	080
саңырау құлақтан	ЗТ/	021	032	041	050	060	074	083
жаным-ай	ЗТ/	021	032	040	051	060	070	082
жүрісіме	ЗТ/	021	032	042	050	060	071	083

Белгілі бір мәтін сөзтұлғаларын белгі-код түріне ауыстырып барып, жиілік сөздіктер түзуге болады немесе әріп күйіндегі мәтіннің өзінен түзілген әліпби-жиілік, жиілік не кері-әліпби сөздіктердегі сөзтұлғаларды (бірліктерді) белгі-код түріне ауыстырып жазуға да болады. Егер ондай сөздіктерді тиісті компьютерлік бағдарлама бойынша «қысылған» түрге

келтіріп жазатын болсақ, зат есім (не басқа сөз табы) сөзтұлғалардың «сипаттық жиілік сөздігін» алуға болады. Бұндай жиілік сөздік қазақ тілінің морфологиялық құрылымын зерттеу нысаны ретінде және әрі қарайғы қазақ тілінің мәтіндер корпусына қажетті маңызды материал болатыны сөзсіз.

Әдебиет

1. *Джубанов А.Х.* Квантитативная структура казахского текста (опыт лингвистического анализа на ЭВМ). Алма-Ата, 1987. С. 147.

2. *Головин Б.Н.* Язык и статистика. М., 1971. 191 с.

3. *Жұбанов А.Қ.* Қолданбалы лингвистика: қазақ тілінің статистикасы. Оқу құралы. – 2 басылым. –Алматы, «Қазақ университеті», 2007, – 209 бет (13,06 б/т).

ҚАЗАҚ ТІЛІНІҢ АВТОМАТТАНҒАН КОМПЬЮТЕРЛІК БАЗАСЫН ҚҰРУ – ЗАМАН ТАЛАБЫ

Мемлекеттік тілді дамыту институты. http://www.mtdi.kz/kz/sybat.
Сұхбатты жүргізген тілші – Мөлдір Мамырбай. Алматы.
16.02.2011, 14:42

Сауал: *Асқар аға, ең алдымен ғылымға келу жолыңыздан бастасаңыз…*

Жауап: Мен Алматыдағы Қазақтың Абай атындағы педагогикалық институттың физика-математика факультетін 1960 жылы үздік бітірген болатынмын. Мамандығым – физика-математика пәндерінің мұғалімі. Сол жылдан бастап, аталған институттың жолдамасымен Қазақтың Мемлекеттік ауылшаруашылық институтының Жоғары математика кафедрасында мұғалім-ассистент болып қызмет істедім. Осы институтта 7 жылдай жоғары математика пәнінен дәріс оқыдым. Сол кездерде Қазақтың Ғылым Академиясының Тіл білімі институтында қызмет істейтін Есет Жұбанов деген ағам Мәскеу, Ленинградтан келген белгілі ғалымдардың тіл білімінің жаңа саласы – математикалық лингвистикадан оқыған дәрістеріне қатысуымды сұрады. Міне, сол дәрістерді тыңдағаннан кейін, математик мамандардың да тіл саласын зерттеуге аса қажет екендігі туралы жақсы түсінген болатынмын. Осы саланың белгілі қазақ ғалымы Қалдыбай Бектаев, институт директоры академик Ісмет Кенесбаев, Бүкілодақтық «Тіл статистикасы» тобының ғылыми басшысы Р.Г. Пиотровский, тілші-ғалым ағайым Есет Жұбанов менің осы жаңа саланы таңдап, аспирантураға түскенімді жөн көріп, кеңес берді. Нәтижесінде, жоғары оқу орнындағы мұғалімдігімді қойып, математикалық тіл білімі деген жаңа мамандық бойынша аспирантураға түстім. Белоруссия астанасы Минск қаласындағы Лингвистикалық университетінде іс-сапармен үш жыл болып, осы саланың белгілі ғалымдарынан дәріс алдым.

Аспирантурада оқып жүріп, М. Әуезовтің «Абай жолы» романындағы тілдік бірліктердің статистикасын және олардың ықтималдықтар теориясының қандай заңдылықтарына бағынатындығын зерттеп, 1971 жылы «ЭЕМ көмегімен қазақ мәтінін статистика-лингвистикалық зерттеу» атты кандидаттық диссертацияны үздік қорғап шықтым. Сол кездегі Қазақ ҒА-ның Тіл білімі институтының директоры Ісмет Кеңесбаевтың қолдауымен институтта қазақ тілі бойынша статистикалық зерттеулерді ЭЕМ көмегімен жүргізетін шағын құрамдағы ғылыми топты алғашында Қ. Бектаев, сосын мен өзім ғылыми жетекшілік еттім. Нәтижесінде, А. Байтұрсынұлы атындағы Тіл білімі институтында сол кезде «Қолданбалы тіл білімі» (қазіргі кезде «Компьютерлік лингвистика») деп аталатын бөлім ашылып, ғылыми зерттеулердің сапасы айтарлықтай арта түсті. Осылайша мен өзімді, әкем – қазақ филологиясының тұңғыш профессоры Құдайберген Қуанұлы Жұбановтың тіл біліміндегі ғылыми жолын жалғастырып келемін деген ойдамын.

Кандидаттық диссертацияма ғылыми жетекшілік еткен ұстазым – физика-математика ғылымдарының кандидаты, филология ғылымдарының докторы – белгілі ғалым Қалдыбай Бектайұлы Бектаев. Қазақстанда математикалық лингвистика саласының дамуына және менің қол жеткізген ғылыми жетістіктеріме ғалым Қ.Б. Бектаевтың сіңірген еңбегі аса зор деуге болады.

2001 жылы «Қазақ мәтінінің мазмұнын формализациялаудың негізгі принциптері» деген тақырыпта докторлық диссертацияны үздік қорғап шықтым. Ғылыми кеңесшім – Минск қаласындағы Лингвистикалық университетінің профессоры, математикалық лингвистика саласының белгілі ғалымы – Александр Васильевич Зубов.

Қорыта айтқанда, қазақ тілін жаңа бағыттардың әдіс-тәсілдерімен зерттеу, негізінен екі кезеңнен тұрады деуге болады, оның алғашқысы – «Қазақ тілінің статистикасы» бағытына байланысты болса, ал екіншісі – «Қазақ тілінің лексикография саласын автоматтандыру» мәселесіне тікелей қатысы бар. Бірінші кезеңде қазақ тілінің әртүрлі мәтіндері (стильдері) бойынша компьютер көмегімен түзілген жиілік сөздіктер негізінде неше түрлі статистикалық мәліметтер алынды, нәтижесінде бірнеше кандидаттық және докторлық диссертациялар қорғалды. 1971 жылдан бері А. Байтұрсынұлы атындағы Тіл білімі институтында (менің тікелей ғылыми жетекшілігіммен және қатысуыммен) «Қазақ тілінің статистикасы», «Компьютерлік лингвистика» және «Қазақ тілі деректерінің компьютерлік базасын құрастыру» мәселелеріне қатысты ғылыми ізденістер орын алып келеді.

Сауал: *Автоматтандыру әдісін толығырақ түсіндіріп беріңізші.*

Жауап: Автоматтандыру мәселесін мынадай жайттан бастағанды жөн көріп отырмын. Мәселен, А. Байтұрсынұлы атындағы Тіл білімі институтының сөз қоржынында көп жылдардан бері жинақталған 5 миллиондай қағаз кәртішкі түріндегі картотекалық қор бар. Осы қордың негізінде Тіл білімі институтында қазақ лексикографтары, кезінде, екі

томдық «Қазақ тілінің түсіндірме сөздігін», бір томдық (түсіндірме) «Абай тілі сөздігін», он томдық «Қазақ тілінің түсіндірме сөздіктерін» құрастырғандары бәрімізге мәлім. Бірақ осындай көлемді тілдік қордан сөздердің мағынасын ашуға қажетті мәліметтерді іздеп табудың оңайға түспегені айтпаса да белгілі. Сонымен бірге аталған қорды толықтыра түсу немесе оны сұрыптау, ұзақ мерзімге сақтау, қайталанатын «кәртішкелерді» қысқарту немесе сапа жағынан жаңарту мәселелерін «қолмен», яғни автоматтандырмай орындау көп күшті және көп уақытты қажет етеді. Аталған (және аталмаған да) әрекеттерді жеңілдету тек қана автоматтандыру әдіс-тәсілдері арқылы, яғни «Қазақ тілінің автоматтанған компьютерлік базасын құру» арқылы ғана жүзеге асады демекпін. Сол сияқты, қолымызда бар лексикографиялық еңбектердің өзін де ұтымды пайдалану немесе олардың өңделген, толықтырылған нұсқаларын баспадан қайта шығару әрекеттері де компьютер арқылы автоматтандыруды қажет етеді.

Бұл айтылғандарға байланысты А. Байтұрсынұлы атындағы Тіл білімі институтында біршама ғылыми ізденістер орын алды. Мәселен, «Абай тілі сөздігі», қазақ тілінің он томдық түсіндірме сөздігі негізінде құрастырылған «Қазақ тілінің сөздігі», жаңадан құрастырылып жатқан 15 томдық «Қазақ тілінің әдеби тілінің сөздігі», екі тілді (қазақша-орысша, орысша-қазақша) 31 томдық салалық терминологиялық сөздіктер, қазақ тілінің 11 томдық ғылыми-техникалық саласына қатысты құрастырылған терминологиялық түсіндірме сөздіктері және т.б. бойынша «Қазақ лексикография саласының автоматтандырылған терминологиялық компьютерлік базасы» іске қосылды.

2009 жылдан бастап біздің шағын ғылыми топ әлемдік деңгейде аса маңызды тақырып – «Мәдени құндылықтар ретіндегі қазақ тіліндегі мәтіндер корпусы және сөздіктердің «Тіл – қазына» атты ұлттық компьютерлік қоры» бойынша ғылыми-зерттеу жұмысы жүргізіле бастады. Мұндай мәтіндік корпусты құрастырудың негізгі мақсаты – аса көлемді мәтін корпусы арқылы айқын контекст бойынша тілдік деректерді тұлғалы етіп көрсету. Сонымен бірге бір рет қана жасалған мәтіндік корпусты көптеген лингвистикалық мәселелерді шешу үшін және қажеттілікке қарай, оны сан рет пайдалана беруге мүмкіндік тудыру.

Қазақ тілінің мәдени құндылығы болып саналатын толық мәтіндер корпустарының мәтіндеріне қажетті грамматикалық белгі-кодтар енгізіледі. Толық мәтіндердің компьютерлік қорының нысандары ретінде Абай Құнанбаевтың, Мұхтар Әуезовтің, Әбіш Кекілбаевтың, Мұқағали Мақатаевтың, Мұхтар Мағауиннің және т.б. толық шығармалар жинақтарынан таңдама мәтіндер алынуда. «Адам-компьютер-адам» жүйесі ретінде қызмет ететін бұл автоматтандырылған «Тіл – қазына» жүйесін әлемдік дәрежедегі ізденістерге сай орындауды мақсат етіп отырмыз.

Аталған жүйе болашақта «Қазақ тілінің ұлттық мәтіндер корпусы» деп аталатын компьютерлік жүйе түрінде көрініс табады деген ойдамыз.

Сауал: *Сіз қазақ тіл білімінің негізін салушылардың бірі, 1937-жылғы сталиндік қуғын-сүргін құрбаны болған профессор Құдайберген Қуанұлы Жұбановтың ұлысыз. Әкеңізді НКВД-ның адамдары тұтқынға алып*

кеткенде, сіздің жарық дүниені көргеніңізге бар-жоғы 6-ақ күн болған деседі. Осыған қатысты не айтар едіңіз.

Жауап: Ол рас, мен 1937 жылдың 13-қарашасында туылсам, әкемізді 19-қарашада «халық жауы» деген жалған айыппен ұстап әкетті. Әкем мені тек перзентханадан алып келіп қана үлгерген екен. Керек десеңіз, әкем менің атымның Асқар екендігін де біле алмай кетіпті.

Әкем Құдайберген Қуанұлы Жұбановтың білімі толысып, күш-жігері кемеліне келіп тұрған шағында, яғни 1938 жылдың 25-ші ақпаны күні, жасы 38-ге жаңа толған шағында атылып кетеді. Тек содан жиырма жыл өткен соң ғана, 1957 жылдың қазан айында әкем – Құдайберген Жұбанов Кеңестер Одағы Жоғарғы Сотының Соғыс алқасының үкімімен біржола ақталып, азаматтық құқығы қалпына келтірілді.

Менің әкем – проф. Қ. Жұбанов өзінің аз жасаған ғұмырында болашақ ұрпаққа зор ғылыми мұра қалдырды. Бірақ, оның бізге келіп жеткені тек біразы ғана. Яғни, қалдырғаны көп болса да жеткені аз. Проф. Қ. Жұбановтың еңбектері мазмұны жағынан алуан түрлі тақырыптарды қамтиды: өйткені ол ең алдымен қазақ тілі грамматикасының, фонетикасының проблемаларын зерттеген ғалым; сондай-ақ қазақ әліппесі, емлесі және терминологиясы мәселелерін шешуге де белсене ат салысқан. Мұнан басқа мәдениет пен өнер жайлы да ауқымды пікір өрбіткен. Оның тіл табиғатынан ізденген нысандары тек фонетика, лексика, морфология, синтаксис мәселелерімен ғана шектеліп қоймай, қазіргі тіл біліміндегі жаңа бағыт деп жүрген мәселелер де әкеміздің сол кездердегі ізденістерінде айтарлықтай орын алған. Яғни ғалымның кейбір ғылыми болжамдары қазіргі уақытта жаңадан ғана қолға алынып жатыр десек те болады. Мәселен, қазір жиі айтылып жүрген «Функционалдық грамматика», «Қолданбалы лингвистика», «Статистикалық лигвистика» деп жүргендерімізді менің әкем – Құдайберген Жұбанов сол кездің өзінде-ақ күн тәртібіне қоя біледі. Бір ғажабы – ғұлама ғалымның топшылаулары мен теориялық тұжырымдамалары әрі нәрлі, әрі терең. Ой-пікірлері соны да ауқымды, жан-жақты болып келеді.

Сауал: *Сол замандарда «ұлтшыл» деп айып тағылғандардың ұрпақтарына да зәбір көрсету әрекеттері көптеп ұшырасты. Бүгінде именбей айтуға хақылы қандай қиындықтарды бастан кешірдіңіз?*

Жауап: Мен 1955 жылы Солтүстік Қазақстан облысының орталығы Петропавл қаласында он жылдық қазақ мектебін бітіріп, Алматыдағы Тау-кен институтына құжатымды тапсырдым. Сол кезде ұл балалардың аса құмартып оқығысы келетіні осы жоғары оқу орны болатын. Қабылдау емтихандарын үздік тапсырып, біраз уақыт студенттер тізімінде де тұрғаным есімде. Бірнеше күн өтісімен ректормен кездесу болып, онда бірнеше талапкерді «халық жауының балалары» деген айыппен оқуға қабылданбағандығы туралы айтты. Сол тізімнің ішінде менің де аты-жөнім жүр екен. Мен үшін бұл хабар қатты күйінуге әкеп соқты. Қалайда институтқа түсесің деп ректордың өзімен сөйлескен әкемнің інісі – Қазақстанға белгілі композитор Ахмет Жұбанов та ешнәрсе істей алмады.

Бірақ сол кезде «Қазақстан мұғалімі» газетінде қызмет істейтін өзімнен үлкен ағам – Есет Жұбанов мені сол кездегі ҚазПИ-дің ректоры – Мәлік Ғабдуллинмен кездестіруге алып барды. Құдайберген Жұбановты тек ірі ғалым ретінде ғана білетін Мәкең, ешқандай басқа сөзге келмей-ақ, әділ шешімге келіп, мені ҚазПИ-дің физика-математика факультетінің студенті етіп бірден қабылдап алды. Содан мен, 1960 жылы ҚазПИ-ді бітіріп, математика және физика пәндерінің мұғалімі болып шыға келдім.

Әкемізді 1937 жылы «халық жауы» деп ұстап әкеткеннен бастап, 1957 жылы ақтағанға дейін анамыздың да осындай халге ұшырау қауіпі туған соң, анамыз 6 баласымен (үлкені – 12 жаста, ең кішісі мен – 1 жаста), Алматыдағы жылы жерімізді суытып, әкеміздің, ата-бабамыздың туған топырағы Ақтөбе облысындағы «Жаңа тұрмыс» аулына қонысымызды аудардық. Соғыс кезі мен соғыстан кейінгі кездердің ең қиын жақтарын алты баласын асырап отырған – Раушан анамыздың көргені көпшілік қауымға мәлім. Үлкен аға-апаларым ер жетіп, мектеп-институттарын бітіріп, жұмысқа тұрып, ел қатарлы коммунистік партия қатарына түскісі келген кездерде де «халық жауының» балаларына қиыншылықтар туды. Кеңес армия қатарында жүрген кезімде, алты баланың ішінен, тек жалғыз мен ғана коммунистер қатарына өттім.

Сауал: *Сұхбат басында жоғары оқу орнында математикадан да дәріс оқығаныңызды айтып қалдыңыз. Тіл біліміне біржолата ауысып, жылдар бойы тұрақтауыңызға не себеп? Бұған әлде әкеңіздің ғұмыр жолы әсер етті ме екен?*

Жауап: Бұл сауалға сөз басында (1-сауалда) математик мамандардың тіл зерттеу саласына қажеттігінен болу керек, - деп айтқан болатынмын. Сонда да, ең алдымен, менің әке жолын таңдауыма себеп болған, әкемнің тіл біліміне соншама еңбек сіңіріп, бірақ дүниеден өте ерте кеткендігі болуы керек және осылай болатынын өзі де білгендей, әкем аз мезгіл ішінде көп тілді үйреніп, адам қайран қалатындай қазақ тіліне қатысты көптеген мәселелердің басын ашып, соңынан соншама мұра қалдырғанына таң қалуымнан да шығар деймін. Өз басым, дәл әкемдей бола алмасам да, ана тілімді мен де қолымнан келгенше зерттеуге әрекеттенсем, о дүниедегі әкем, бір аздаған болса да, риза болар-ау деген сезімнен туындауы да мүмкін.

Тіл біліміне біржолата ауысып, тұрақтауымның тағы бір негізгі себебі – «Әкең өлсе де, әкеңді көргендер өлмесін» деген қазақ мақалы болар деп санаймын. Яғни мен, өз басым, әкемді көрмей өссем де, әкемді көргендердің, әкемді ғалым ретінде жақсы білетіндердің және әкем жайлы сөз болатын орта арасында қызмет етуді мақсат еткен сияқтымын. Мәселен, мен Тіл білімі институтында аспирантура курсын тамамдап, ғылыми қызметкер міндетін атқарып жүріп, әкемді көздері көрген Тіл білімі институтының директоры, ҚазССР ҒА академигі Ісмет Кеңесбайұлы Кеңесбаевтың, әкемнің шәкірттері филология ғылымдарының докторлары, профессорлар, ҚазССР ҒА корреспондент мүшелері Мәулен Балақайұлы Балақаев пен Ғайнетдин Ғалиұлы Мұсабаев, әкемнің өмірі мен шығармашылығын зерттеген академик Рәбиға Ғалиқызы Сыздықтармен бір ұжымда қызмет істегенімді мақтаныш

етемін. Осындай ірі ғалымдардың аузынан әкем жайлы жылы лебіздерді естігенде, шынында да, әкемді көргендей болатыным рас. Сол сияқты, Институттың ғылыми кеңесінің мүшесі болып жүріп, қазақ тілінен кандидаттық, докторлық диссертациялар қорғау кезінде менің әкемнің аты аталмайтын кезі болмайды деуге болады. Ғылыми кеңес кезінде әкемнің атын әр естіген сайын, өзімді математика саласынан қазақ тіл білімінің қыры мен сырын зерттеуге арнап жүргеніме тек қана риза боламын.

Қорыта айтқанда, негізгі мамандығым математика саласы бола тұра, мен тіл білімін саналы түрде таңдап алдым деуге әбден болады.

Сауал: *Марат Қабанбай марқұм бір Бейімбетке мақала жазу үшін оның шығармаларын 24 рет оқып шығыпты. Сіздің еңбектеріңіз болса, «Абай жолы» роман-эпопеясымен тікелей байланысты. Осы іргелі туындыны осы уақытқа дейін қанша рет оқып шықтыңыз?*

Жауап: Меніңше, шығарманы дұрыс түсіну үшін қайта-қайта оқи берудің қажеті де жоқ шығар. Себебі ол адамның қабылдау қабілетіне тікелей байланысты ғой. Ал біздің зерттеудің алға қойған мақсаты мүлдем басқа, ол шығарманы оқып шығу санымен өлшенбейді.

Мен, төрт томнан тұратын «Абай жолы» романы мәтінінің бірінші сөзінен соңғы сөзіне дейін, олардың қай сөз табына қатыстылығына қарай шартты белгілер қойып, талдап шыққан адаммын.

«Абай жолы» романын мектеп қабырғасында жүрген кезімде оқығаным есімде. Кейіннен кемелденген көзқараспен тағы да оқып шықтым. «Абай жолы» романының стильдік ерекшелігін білу үшін тек мазмұнын түсініп қою ғана емес, автор әрбір сөзді қалайша пайдаланды, қандай сөз тіркестерін қолданды, қандай сөздер жиі кездесті, ал қайсылары сирек кездесті деген мақсатта зерттеулер жүргізілді. Тіпті, бір беттегі әр сөйлемді бірнеше рет оқуға тура келген кездер де болды. Неше түрлі омонимдік, синонимдік, кірме сөздерді ажырататын шартты белгілер қойып, олардың статистикалық сипатын анықтадық. Біздің бастамамызбен «Абай жолының» электронды түрі, жиілікті, әліпбилі, кері әліпбилі сөздіктері ЭЕМ-де құрастырылып, нәтижесінде бірнеше кандидаттық және докторлық диссертациялар қорғалды. Сондықтан біз «Абай жолы» романын статистикалық тұрғыда айтарлықтай жақсы зерттедік деуімізге негіз бар.

М. Әуезовтің тек «Абай жолы» романы ғана емес, біз автордың 20 томдық шығармалар жинағының да жиілік сөздіктерін ЭЕМ арқылы алып, тиісті статистикалық зерттеулер жүргізіп жатырмыз. Бірақ оларды бірнеше рет оқып шықтық деп айта алмаймыз және ол біздің мақсатымыз да емес.

Сауал: *««Жазушы» баспасынан шыққан «Абай жолында» қате деген өріп жүр» деген пікірге қосыласыз ба?*

Жауап: «Өріп жүр» деу, әрине, артықтау, бірақ қателердің аздап кездесетіні рас. Оларды шығарушы баспаның жіберген қателері деп қабылдау керек. Қатені жібермеу жағынан қазір компьютер мүмкіндігін молынан пайдаланғанды жөн санар едім. Болашақта «Абай жолы» романының мәтіні немесе басқа да шығармалар мәтіндері біздің құрастырып

жатқан қазақ мәтіндерінің корпустық компьютерлік базасына енгізілген жағдайында ондай қателер орын алмауы керек деп түсінемін.

Сауал: *Қазіргі сөздіктердің сапасына сіздің пікіріңізді білгіміз келеді. Аударма саласына көңіліңіз тола ма?*

Жауап: Сөздіктердің түрі көп қой, сол себепті мен барлық сөздіктердің сапасы жайлы айта алмаспын. Сондықтан, тек ғылыми-техникалық мәтіндерді қазақшаға аударғанда кездесетін кейбір кемшіліктерді ғана сөз етейін. Жоғарыда айтқанымдай, екі тілді салалық терминологиялық сөздіктер, әр түрлі ғылыми-техникалық 31 сала бойынша қазақша-орысша және орысша-қазақша сөздіктер (және тағы да басқа) терминдерге қатысты аударма сөздіктердің баспадан шыққаны белгілі. Бір жағынан олар бекіту, стандарттау комиссиясының алдынан да өтіп, «бекітілген терминдер» сөздіктері қатарына жатады. Негізінен, орыс тілінен қазақшаға аударылған мәтіндерде, әлде де, көптеген терминдердің стандартталды, бекітілді дегеннің өзінде, олардың басқаша аударылған бір емес, бірнеше вариантта қолданылып жүрген көріністері жиі кездеседі. Сондықтан, әр ғылыми салаға тән терминдер бір ізге түсе қойды деуге және жарық көрген терминологиялық сөздіктерге көңіл тояды деп айтуға әлі ерте тәрізді. Ғылыми-техникалық терминдерді қазақшаға аудару мәселесі әлде де көп өңдеуді, басқаша көзқарасты қажет етеді. Әсіресе, бір тілден екінші тілге аударған кезде, мысалы, орыс тіліндегі немесе орыс тілі арқылы енген халықаралық терминдердің барлығын өз тілімізге аударуға тырысып, кейде сөз мағынасын шатастырып алатын тұстар да жиі кездесіп тұрады. Сол себепті, Құдайберген Жұбановтың 10 принципіне сүйене отырып, кейбір жағдайларда Халықаралық терминдерді сол күйінде қалдырғанды, өз басым, дұрыс деп санаймын. Қысқаша айтқанда, қазақ тіліне қатысты «Аударма теориясы» әлде де арнайы зерттеулерді қажет етеді демекпін.

Сауал: *Осы жылдан бастап дәстүрлі кандидаттық және докторлық диссертациялар қорғайтын ғылыми кеңестер жабылып, кейбір шетелдегілерше магистрлік, PhD докторлық деңгейде қорғайтын ғылыми кеңестер жоғары оқу орындарында көбірек ашылмақшы. Бұл процеске қандай баға бересіз?*

Жауап: ҚР БжҒМ Абай атындағы Қазақ Ұлттық педагогикалық университеті жанындағы Магистратура және PhD докторантура институтындаөз саламнандәріс оқып келе жатқанымма төрт жылдай уақыт болды. Ал ҚР БжҒМ ҒА А. Байтұрсынұлы атындағы Тіл білімі институтының кандидаттық, докторлық қорғайтын ғылыми кеңесінің мүшесі болғаныма он жылдай уақыт өтті. Сондықтан кандидаттық, докторлық қорғайтын ғылыми кеңесті тоқтатып, тек қана магистрлік, PhD докторлық қорғау ғылыми кеңестерін ғана ашуды мен дұрыс деп санамаймын. Бұл екеуі қатар жүруі керек. Мысалы, Ресейде кандидаттық, докторлықты қорғау ғылыми кеңестері өз жұмыстарын жалғастыруда. Басқаларды айтпағанның өзінде, менің қаншама шәкірттерім кандидаттық және докторлық жұмыстарын дайындап, аяқтауға жақын қалғанда, оларды қорғатпай тастатуды дұрыс деп санамаймын. Бірақ әлде де үміт үзбей, министрлікте

оларға бір амалын қарастырар-ау деген ойдамын. Ал үміткерлерді ғылыми кеңесте сапалы қорғату – басқа мәселе. Оны сөз етпей-ақ қояйын.

Сауал:*Тіл білімі институтында кандидаттық, докторлық қорғайтын ғылыми кеңестің мүшесісіз. Ел президенті Н.Назарбаевтың халыққа жолдауында сөз болған, 60 қорғалған ғылым кандидатының тек екеуі ғана ғылым саласын таңдайды деген сөзіне сіз қалай қарайсыз, жалпы сапалы және сапасыз қорғаулар сіздің ғылыми кеңесте бола ма? Статистикалық және компьютерлік лингвистика салалары бойынша қанша адам қорғап үлгерді?*

Жауап: Менің тұжырымдауымда, ондай мәліметті президентімізге дайындап берген министр ме, әлде басқа ма білмеймін, бірақ олай жобалап, негізсіз айту дұрыс емес дер едім. Ғылымдағы бетке ұстар жаңалықтар қорғалған диссертация санымен есептелмеуі керек деп санаймын. Анда-санда болатын ғылыми жаңалықтың өзі барлық қорғалған диссертациялар шығынын ақтайтынына толық сенем. Жылына кем дегенде 10-20 кандидаттық диссертация, 5-6 докторлық диссертация қорғалады. Олардың ішінде жақсылары да, ал кейде төмен деңгейлердегі де кездеседі. Тағы да қайталасам, ғылыми жаңалықтар, оның даму қарқыны санақпен айқындалмайды. Қорғалған диссертациялардың бәрінде бірдей жаңалық бола бермейтіні рас, бірақ оның ішіндегі бірең-саран жаңалықтардың өзінен біз ғылым дамуына қатысты көп нәрсе ұтатынымыз айқын. Әрине, бұл менің субъективті пікірім.

Енді сауалдың екінші бөлімі – статистикалық және компьютерлік тіл білімі саласына қатысты қорғалатын диссертациялар жайлы айтсақ, олар жаңа бағыт болғандықтан ба, тілшілердің есеп-қисапқа қатысты білімінің аздығынан ба, олардың саны 2-3-тен аспайды. Оның негізгі себебі филологтардың жоғары оқу орындарында мұндай пәндерді өтпеуінен деуге болады. Білім және ғылым министрлігі, ғылым комитеті осы мәселеге үлкен мән берулері керек деп білемін.

Сауал:*Қазақ тілінің қоржынында дәл қазір қанша сөз бар?*

Жауап: Әзірше шамамен ғана айтпаса, бұл сауалға дәл жауап беру қиынға соғады. Мәселен, «Абай тілі сөздігінде» Абайдың сөз байлығы 6 мыңдай дара сөз бен 1 мыңнан аса сөз тіркестері бар екен. Екі томдық «Қазақ тілінің түсіндірме сөздігінде» 18135 сөз, 3371 фразеологиялық тіркес қамтылған, 10 томдық академиялық түсіндірме сөздікте 67 мыңдай атау сөз, 24 мыңнан аса фразеологиялық тіркес қамтылған. Ал жаңадан құрастырылып жатқан 15 томдық «Қазақ әдеби тілінің сөздігінде» кірме сөздерді, көнерген сөздерді, жаңа сөздерді де қоса есептегенде, 100 мыңға жуық сөз бен сөз тіркестері жинақталады деп үміттенеміз. «Қазақ әдеби тілінің сөздігіндегі» сөздердің дәл санын тек 15 том толығымен баспадан шыққан күні ғана біле алатын шығар. Кейбір шалағалымдар «қазақ тілін байытамыз» деген оймен, аударуға келетіндіні де, келмейтіндіні де қазақшалап, қазақ сөзінің қатарына қоса беру дұрыс деп санамаймын. Қазақтың төл тіл байлығы 80 мыңнан асқанның өзі де қазақ тілі үшін – үлкен қазына.

Сауал: *«Қолданбалы лингвистика: қазақ тілінің статистикасы» атты бір ғана еңбегіңізді даралап алып үңілетін болсақ, ЖОО қабырғасындағы студентке аталған оқу құралын ұғу қиынға соғады деп есептемейсіз бе?*

Жауап: Жоқ. Оқу құралын қазіргі қазақ филолог-студенттің деңгейімен жазуға болмайды деп санаймын. Ол әлемдік деңгейдегі нұсқаны ұстаным етіп жазылуы керек. Менің авторлығыммен жарық көрген «Қолданбалы лингвистика: қазақ тілінің статистикасы» және т.б.меңгеру үшін студентке математика пәнінен мектеп қабырғасынан алған білімі жеткілікті. Сондықтан, студенттің түсіну-түсінбеуі бір бөлек те, ғылыми талап бір басқа. Мысалы үшін, мен 5 жыл бойы әл-Фараби атындағы Ұлттық университетінде «Қолданбалы тіл білімі» мамандығы бойынша дәріс оқыдым. Сол жердің оқыту талабы бойынша 3 оқу құралын жаздым. Ең бірінші «Қолданбалы лингвистика: Қазақ тілінің статистикасы» атты оқу құралым жарық көрді. Екінші «Қолданбалы лингвистика: формалды модельдер» деген, соңғысы «Компьютерлік лингвистикаға кіріспе» деген оқу құралым шықты. Абай атындағы мемлекеттік университеттен бөлініп, қазір жеке отау тіккен магистранттар, PhD докторанттар дайындайтын институтта да аталған оқу құралдары ғылымилығы мен құндылығын жойған жоқ. Алған тәжірибем бойынша, аталған оқу құралдарындағы материалдарды филолог-студенттер түсінуге қиын екен деп айта қоймады. Студенттер немесе магистранттар оқу құралындағы материалдарды түсінуге талпынып, тырысулары қажет деп санаймын. Көбінде, тіл білімін таңдап келген студенттердің өздері математикадан ат-тондарын ала қашқысы келіп тұратыны белгілі ғой. Сондықтан ЖОО қабырғасындағы студентке аталған оқу құралдары ұғыну жағынан қиынға соғады деп есептемеймін.

Сауал: *Осы уақытқа дейін бас-аяғы қанша кітаптың авторы атандыңыз?*

Жауап: Кандидаттық, докторлық диссертацияларым және басқа да ғылыми ізденістерім монография түрінде, қолданбалы лингвистика саласы бойынша ҚР әл-Фараби атындағы Қазақ Ұлттық университетінде «Қолданбалы тіл білімі» мамандығы бойынша оқыған дәрістерім 3 оқу құралдары түрінде жарық көрді. «М.О. Әуезовтің «Абай жолы» романының жиілік сөздігі», «М.О. Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері», «Куманша-қазақша жиілік сөздік» атты сөздіктер авторларының бірі болғанымды есепке алғанда, оннан аса монографиялар мен сөздіктер баспадан шықты. 70 жылдық мерейтойымның қарсаңында менің таңдамалы ғылыми шығармаларым «Қолданбалы тіл білімінің мәселелері» деген атпен «Қазақ ғылымының озық үлгілері» мегажобасы аясында және ҚР Мәдениет және Ақпарат министрлігі, Тіл комитеті мен ҚР БҒМ А. Байтұрсынұлы атындағы Тіл білімі институты Ғалымдар кеңесінің ұсынысымен 60 баспа табақ көлемінде, 2000 таралыммен басылып шықты. Ғылыми мақалалардың саны 100-ден асады. Құдай қаласа, әлде де жаңадан зерттеліп жатқан ғылыми ізденістерім монография түрінде, ал магистранттар мен PhD докторанттарға оқып жатқан дәрістерім оқу құралдары ретінде жарық көреді деген үміттемін.

Сауал: *Өзіңіз жаңалық алып келген тіл біліміндегі жаңа бағытты дамытуға үлес қосып жүрген шәкірттеріңіз бар ма?*

Жауап: Жақын шетел мен Қазақстанның әр облыстарынан дайындаған шәкірттерімді қоса есептегенде, қазақ тілінің қолданбалы саласы бойынша 8 ғылым кандидатын дайындап шығардым. Ұстазымыз Қалдыбай Бектайұлының ғылыми мектебінің орталығы Шымкент қаласы болғандықтан болу керек, тіл статистикасына қызығушылық танытатын, негізінен, осы қаланың жастары деуге болады. Сол сияқты, өзім 6 жыл қызмет істеген әл-Фараби атындағы Қазақ Ұлттық университетінің «Қолданбалы тіл білімі» мамандығы бойынша оқыған дарынды жастар да А. Байтұрсынұлы атындағы Тіл білімі институтының «Компьютерлік лингвистика» бөлімінде жаңа саланы меңгеріп, ғылыми дәреже алуға барынша талаптануда.

Сауал: *Бүгінде Тіл саясатын дамытуға көп қаржы құйылып жатыр. Жасыратын несі бар, сол қаражат ауаға ұшып жатқандай...*

Жауап: Қаражаттың орынды жұмсалып жатқанына менің де күмәнім бар. Кей кезде аты бар да, заты жоқ дегендей, жоғары жақтан ғылымға «мұнша қаржы бөлдік» деп ауыз толтырып айтылғанымен, ол қаржының қайда кетіп жатқаны белгісіз. Мысалы, біздің саланың «Қолданбалы» деген атының өзі қажеттілікті аңғартса да, бұл салаға министрлік және Ғылым комитеті жағынан тиісті көңіл аударылмай келе жатқаны аян. Мәселен, әл-Фараби атындағы Қазақ Ұлттық университетінде «Қолданбалы тіл білімі» мамандығын әуелі ашып, 6 жылдан кейін оны жауып тастауына қандай себеп болғаны тіпті де беймағлұм. Сол сияқты, А. Байтұрсынұлы атындағы Тіл білімі институтында «Қолданбалы тіл білімі» бөлімін жауып тастап, оның орнына заңдастырылмаған шағын ғана «Компьютерлік лингвистика» ғылыми тобы аса қиындықпен жоспарлы тақырыптар бойынша зерттеулер жүргізіп келеді.

Сауал: *Қазақ тілі өзінің «мемлекеттік тіл» рөлін атқара алып жүр ме?*

Жауап: Қазақ тілі өзінің «мемлекеттік тіл» рөлін әлі де болса толық атқара алмай келе жатыр дер едім. Оның себебін арнайы зерттеу тақырыбы етіп алу керек сияқты. Мүмкін, мемлекеттің көпұлтты болуынан ба, әлде Үкімет басында басқарып отырған басшылардың осы мәселеге айтарлықтай мән бермей, салғырттық жасауынан ба, әйтеуір, меніңше қазақ тілін, әзірше, «мемлекеттік тіл» деуге болмайды деп санаймын. Әрине, бұрынғымен салыстырғанда қазіргі кезде аздаған болса да ілгерілеу бар. Менің пайымдауымша, «мемлекеттік тіл» дәрежесіне лайықты болу үшін – сол тілде шығатын ғылыми еңбектер саны көбеюі керек. Бізде осы күнге дейін нақты ғылымдар саласы бойынша шығатын ғылыми еңбектер көбіне-көп орыс тілінде жарық көруде. Физика, химия, металлургия саласындағы ғалымдарымыздың қазақша еңбектері көптеп шыға бастаса ғана ана тіліміздің дәрежесі жоғарылайды және оны «мемлекеттік тіл» рөлін атқарады деп айтуға болар еді.

Сауал: *Сіздің еңбектеріңіздің барлығы «автоматтандыру, компьютерлік база құрастыру» жайлы болып келеді. Өзіңіз компьютердің құлағында еркін ойнайтын шығарсыз?*

Компьютерді жақсы білмесек, біздің жұмысымыздың алға баспайтыны айдан анық. Осы саланың тізгінін басынан бастап ұстағаннан соң, ең алдымен, әрине, өзім жақсы білуім керек. Компьютердің «құлағында еркін ойнау» үшін, ең әуелі, кез келген тілге қатысты «есептерді» компьютерге жүктей білетіндей жақсы бағдарламашы (программист) болу керек деп санаймын. Кезінде ол сатыдан жеке басым өтсем де, қазір заманға сай, бұл істі жастарға жүктеу керек деп санаймын. Ал менің міндетімде – туындаған мәселені компьютерде шеше алатын ең ұтымды жолды, программистер тілімен айтқанда, сол есептің ұтымды алгоритмін таңдай білу, соның ұтымды екендігін дәлелдей білу керек дер едім. Сондықтан компьютердің «құлағында еркін ойнау» деген – көп мағыналы сөз тіркесі. Оған компьютердің жеке-жеке құрамдарын жөндейтін мамандар немесе компьютер әрекетін басқаратын жүйелік бағдарламалар қоятын мамандарды да (системщиктерді) жатқызуға болар еді. Олардың жұмыстары өздеріне тән болса, менің ғылыми жұмысым өзіме тән деуге болады.

Сауал: *Соңғы сауал ретінде білгім келгені, осы қазақ тілінің қолданбалы саласының жаңа бағыты бойынша қандай мәселелер негізгі назардан тыс қалып келеді деп санайсыз.*

Жауап: Кеңес одағының құрамында болған басқа ұлттық мемлекеттерге қарағанда Қазақстан – «Қолданбалы тіл білімі» деп аталатын қазақ тіл білімінің жаңа саласы бойынша көш бастап келеді деуге толық негіз бар. Ол, әрине, менің ғылыми ұстазым – ықтималдықтар теориясы мен математикалық статистика пәнінің маманы, физика-математика ғылымдарының кандидаты, филология ғылымдарының докторы, профессор Қалдыбай Бектайұлы Бектаевтың қазақ тіліне сіңірген сүбелі еңбегі деп түсінуіміз қажет. Өзбек, қырғыздарға және т.б. қарағанда көш ілгері тұрғанымызды мақтанышпен айтуымызға болады. Алайда, даңдайсып кетуге де тағы болмайды. Себебі біздің алдымызда істегеніміздан гөрі, істемегеніміз әлі де көп. Неге десеңіз, бұл әлде де жаңа сала және оған тиісті дәрежеде арнайы көңіл бөлінуі керек деп тұжырымдаймын. Қазақ тілінің қолданбалы саласының жаңа бағыттарына тек Тіл білімі институты басшылығы жағынан ғана емес, Тіл комитеті жағынан да, Ғылым комитеті жағынан да тиісті көңіл бөлінуі қажет.

Ал, қазақ тіліне қатысты, әлі де өз шешімін күтіп тұрған мәселелер жеткілікті. Бұл қысқаша сұхбатта оларды тізіп айтпай-ақ, тек кейбіреулеріне ғана қысқаша тоқталуға болады. Қазақ тіліне қатысты, ең алдымен, тілдік бірліктерді автоматты түрде «тану» мәселесі. Морфологиялық деңгейде ол – сөз таптарын автоматты түрде ажырата білу мәселесі болса, синтаксистік деңгейде ол – сөйлем мүшелерін автоматты «тану», ал семантикалық деңгейде лексикалық бірліктердің немесе сөз тіркестерінің, сөйлемдердің, мәтін абзацтарының, толық мәтіндердің автоматты түрде мағынасын, мазмұнын ашу және т.б.

Тағы бір шешімін күттірмейтін маңызды мәселе – машиналық (компьютерлік) аударма мәселесі. Қазіргі кезде үнді-еуропа тілдеріне қатысты «машиналық аударма» мәселесі 70-80 пайызға дейін шешілді деуге болады. Ал, басқа түркі тілдерін былай қойғанда, қазақ тіліне ғана қатысты бұл мәселе бүгінгі таңда барынша кешеуілдеп тұрғаны белгілі. Министрліктен, Ғылым комитеті жағынан бұл мәселеге арнайы назар аударылса екен дегім келеді.

Сөз соңында айтарым, қысқаша түрде атап кеткен бұл жайттар ана тіліміз үшін толғағы жеткен, шапшаң шешу жолдарын қарастыруға тиісті маңызды мәселелер қатарына жатады.

ҚАЗАҚ ТІЛІНІҢ АННОТАЦИЯЛАНҒАН МӘТІНДЕР КОРПУСЫНДАҒЫ СЫН ЕСІМ СӨЗДЕРДІҢ ЛЕКСИКА-МОРФОЛОГИЯЛЫҚ ҚҰРЫЛЫМЫНА БЕЛГІ-КОД ҚОЮ БАҒДАРЛАМАСЫ

ҚР БҒМ ҒК А.Байтұрсынұлы атындағы Тіл білімі Институты. Аса көрнекті тіл маманы, филология ғылымдарының докторы, профессор Қайырболат Есеновтің туғанына 80 жыл толуына арналған «Қ.Есенов және қазақ грамматикасының өзекті мәселелері» атты халықаралық ғылыми-теориялық дөңгелек үтел материалдары. Алматы-2012, 41-49-бб.

Корпустық лингвистика тіл білімінің жеке саласы ретінде өзімен іргелес жатқан тіл ғылымы пәндерімен жанасып жатады, яғни математикалық лингвистика, дискурстік анализ және лексикография салаларымен жақын. Корпустық лингвистиканың басқа тіл ғылымы пәндерімен қарым-қатынаста болу ерекшелігі, бір жағына алғанда, мәтіндер корпусының корпустық лингвистика қызметінің нәтижесі ретінде болса, ал екіншіден – лингвистикалық пәндердің басқа түрлеріне де бастапқы эмпирикалық материал болу мүмкіндігінде. Міне, дәл осы жағдай корпустық лингвистиканың фонетика, лексикология, грамматика және стилистика салаларымен тығыз байланыста болатындығына негіз бола алады.

Грамматика саласындағы сөз табының мазмұны «лексика-грамматикалық» деп аталатын екі компоненттің бірлігінен құралатыны белгілі. Бұл атаудан лексикалық семантикамен бірге грамматикалық семантика деген ұғымды да бірге қабылдау керек болады. Ал қос сөздің екінші бөлігіндегі «грамматикалық» сөзінің мазмұнына белгілі бір сөзге тән грамматикалық категориялардың және олардың жасалу, түрлену тұлғаларының мағыналары енеді. Яғни бұл жердегі «грамматикалық» ұғымына сөздің жаңадан сөз тудыру, сөз түрлендіру, сөз байланыстыру жүйелерінің мағыналары мен тұлғалары, демек, бүкіл сөздің морфологиялық және синтаксистік белгілері (сыр-сипаттары) түгел енеді [1, 126-134-бб.].

Қазақ тілінің қолданбалы саласының, дәлірек айтқанда, статистикалық лингвистиканың алдына қойған мақсаттарының бірі – сөз табының лексикалық (семантикалық), грамматикалық (морфологиялық және синтаксистік) белгілері арқылы, яғни осы үш белгінің негізінде статистикалық зерттеулер жүргізіп, сөздің мағынасын, морфологиялық формасын (тұлғасын) және синтаксистік қызметін сандық деректер арқылы тани білу. Соның нәтижесінде мәтін мазмұнын ашудың формальды жақтарын айқындау.

Қазақ тілінің қолданбалы саласының, дәлірек айтқанда, статистикалық лингвистиканың алдына қойған мақсаттарының бірі – сөз табының лексикалық (семантикалық), грамматикалық (морфологиялық және синтаксистік) белгілері арқылы, яғни осы үш белгінің негізінде статистикалық зерттеулер жүргізіп, сөздің мағынасын, морфологиялық формасын (тұлғасын) және синтаксистік қызметін сандық деректер арқылы тани білу. Соның нәтижесінде мәтін мазмұнын ашудың формальды жақтарын айқындау.

Сөз таптарының ішкі жіктелген бөліктерінің түрлері тілде біреуі аз не көп қолдануы мүмкін. Статистикалық зерттеу тәжірибесі бойынша олардың қолдану жиілігін анықтау үшін кейбір шараларды алдын ала белгілейтін бағдарламаларды іске қосу қажет.

Мысалы, әр түрлі тілдік стильдердегі сөз таптарының қатынастық сипатының статистикасын білу керек болса, әр сөз табына қатысты шартты белгіні (кодты) қоюдың зерттеушіге ыңғайлы жолдарын ойластыруымыз керек. Себебі, сөз таптарын не басқа тілдік бірліктерді мәтін ішінде түр-тұрпатына қарай формальды әдіспен ажырату мүмкіндігі әлі де болса зерттеу аясынан тыс қалып жүр.

Орыс тілінің морфологиялық сипатына тән сөз таптарын шартты түрде белгілеу (кодтау) ресейлік ғалым Б.Н.Головиннің «Язык и статистика» атты кітабында да сөз болады [2].

Ал морфологиялық ақпарат белгілерін сәйкестендіру бағдарламасын құрудағы біз ұсынып отырған үлгінің ерекшелігі – қазақ тілі сөз таптарының ішкі бөліктерінің құрылымдық сипаты мен жасалу жолдары ескеріліп, әмбебаптық сипатты көздейді. Алғашқы зерттеулерімізде қазақ тілінде ең жиі қолданыстағы зат есім мен етістіктік сөздердің лексико-морфологиялық деңгейдегі белгіленім қою мәселелерінің бағдарламасы қарастырылды [3; 4; 5].

Мысалы, зат есіммен, үстеумен және т.б. сөз таптарымен әрі ортақ, әрі тұлғалас сөз тудыратын жұрнақтарымен қатар тек туынды сын есім ғана жасайтын арнаулы жұрнақтары және сын есімнен сын есім тудыратын шырай жұрнақтары да бар. Бұл аталғандар да сын есімнің морфологиялық формалары болып есептеледі.

Сын есімдердің тағы да бір өзіндік қасиеті – заттардың алуан түрлі сыр-сипаттары мен белгілерін тікелей білдірумен қатар басқа заттардың қатынастары арқылы да сондай сипаттарды білдіре алады. Осыған байланысты, сын есімдер семантикалық мағыналары мен грамматикалық

ерекшеліктеріне қарай, *сапалық (негізгі)* және *қатыстық (туынды)* сын есім деп аталатын екі салаға бөлінетіні де белгілі.

Негізгі немесе сапалық сын есімдер жайлы айтатынымыз, олар, әдетте, ешбір қосымшасыз тұрып-ақ, заттың әр қилы сыр-сипатын білдіретін түбір сөздерден тұрады. Бірақ, егер кейбір түбір сөз деп жүрген сөздерді олардың тарихи даму тұрғысынан қарастырсақ, олар туынды сөз болып та шығуы мүмкін. Сондықтан, құрылымы жағынан негізгі сын есім деп есептелетін сөздердің саны да, сапасы да ұдайы өзгеріп отырады. Ал сын есімдерді морфемалық құрылымына қарай, негізгі және туынды деп жіктеу шартты екендігін де есте ұстаған жөн.

Сонымен кез келген сөйлем ішінде (мәтінде) кездесетін сын есім сөздер, біріншіден, сапалық сын есімге жатуы мүмкін *(ақ, қара, сары, көк, сұр, биік, үлкен, аласа, жылы, суық, ірі, кіші* т.т.), екіншіден, морфологиялық (синтетикалық) тәсіл бойынша, яғни тиісті жұрнақтар арқылы жасалған туынды сын есімге жатуы мүмкін екен.

Үшіншіден, қарастырылатын кезекті сын есімге қатысты сөз синтаксистік (аналитикалық) тәсіл, яғни жалаң сын есімдердің бір-бірімен тіркесуі арқылы *(ақ сары, қызыл сары, ақ шұбар, ақ көйлекті, көп балалы, үлкен-кішілі* т.т.) және морфологиялық-синтаксистік (семантикалық) тәсіл бойынша жасалатын туынды сын есімдер де болуы мүмкін.

Соңғы аталған тәсіл бойынша туатын сын есімдер қолданылуы жағынан алғанда тіпті де өнімсіз деуге болады. Сондықтан сын есімдердің жасалу жолдарының ең негізгісі деп, ең алдымен морфологиялықты немесе синтетикалықты, ал одан кейін барып, аналитикалықты, яғни синтаксистік тәсілді айтуға болады.

Осы себептен, біз сын есімнің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасын құрастыруда тек қана сын есімнің синтетикалық тәсілмен жасалу жолына ғана, дәлірек айтқанда, есімдерден сын есім тудыратын өнімді жұрнақтар арқылы жасалу жолын негіз етпекпіз.

1-кесте. **Сын есімнің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы**

Белгі-код орны	**Сын есім сөздердің лексика-морфологиялық сипаты**	**Белгі-код**
1	**2**	**3**
1	**Егер сөз (сөзтұлға) морфемалық құра-мына қарай сапалық сын есім (негізгі сын есім) болса:**	СН1
2	1) заттың не құбылыстың түрі мен түсін анықтаса. Мыс.: *ақ, қара, қызыл, жасыл* т.б.	201
	2) заттың не құбылыстың сыры мен сапасын анықтаса. Мыс.: *жақсы, жаман, тәуір, нашар* т.б.	202
	3) заттың көлемі мен аумағын, ұзындығы мен салмағын анықтаса. Мыс.: *үлкен, кіші, ұзын, ауыр, қысқа, жеңіл* т.б.	203
	4) заттың дәмі мен иісін білдірсе. Мыс.: *ащы, тәтті, күлімсі* т.б.	204
	5) заттың не құбылыстың басқа да қасиет-белгілерін білдірсе.	205

1	**Егер кезекті сөз (сөзтұлға) морфемалық құрамына қарай қатыстық сын есімге жатса, яғни заттың белгісін басқа бір заттың қатысы арқылы білдірсе және ондай туынды сын есімдер өнімді (өнімсіз) жұрнақтар арқылы жасалса.**	СН2
2	Есімдерден туынды сын есім тудыратын өнімді жұрнақтар арқылы жасалып, заттың сыртқы түрі мен түсіне, кескіні мен келбетіне, сыры мен сынына, мекен мен мезгілге және т.б. белгілеріне қатысты сындық ұғым білдірсе. 1. ***-қы, -кі; -ғы, -гі*** жұрнақтары арқылы кейбір зат есімдерден, есімдіктерден, үстеулерден, сондай-ақ, жатыс және шығыс септіктегі сөздерден жасалатын туынды сын есімдер мынадай мағынада жұмсалса: а) егер мекендік ұғым білдіретін кейбір зат есімдерге, үстеулерге, сондай-ақ, жатыс (кейде шығыс) септік формаларындағы есімдерге жалғанса. Мыс.: *ауызғы, төргі, түпкі, ішкі, төменгі, соңғы, санаулы, бүктеулі* т.б.	210
	ә) егер мезгілдік ұғым білдіретін кейбір зат есімдер мен есімдіктерге, мезгілдік үстеулерге жалғанса. Мыс.: *кешкі, түскі, күзгі, жазғы, көктемгі, түнгі, күндізгі* т.б.	211
	2. ***-лы, -лі, -ды, -ді, -ты, -ті*** қосымшасы арқылы мынадай туынды сөздер жасалады: а) белгілі бір заттың (я құбылыстың) бар екендігін білдіру үшін зат есімдерге жалғанса. Мыс.: *арлы, сулы, әсерлі, гүлді, икемді, байыпты, инабатты, балалы* т.б.	212
	ә) қосарланған зат есімнен, сын есімнен, сан есімнен, үстеуден күрделі сын есімдер осы қосымшалар арқылы жасалса. Мыс.: *ағалы-інілі, ойлы-қырлы, таулы-тасты, өзенді-сулы, үлкенді-кішілі, бұрынды-соңды* т.б.	213
	б) екі-үш сөзден құралып, суреттеме атаулар қызметіндегі күрделі сын есімдер жасалса. Мыс.: *ақ басты, қаз мойынды, ай қабақты, теке сақалды* т.б.	214
	3. ***-сыз (-сіз)*** жұрнағы есім сөздерге жалғанып, болымсыздық мағынадағы туынды сын есімдер жасалса. Мыс.: *баласыз, көліксіз, білімсіз, сенсіз, бізсіз, мұңсыз* т. б.	215
	4. ***-шыл (-шіл)*** қосымшасы зат есім, есімдік, модаль сөздерге (әр тарап) жалғанып, бейімділікті, икемділікті, құмарлықты білдіретін қатыстық сын есімдер жасалса. Мыс.: *ұйқышыл, ұйымшыл, өзімшіл, турашыл* т.б.	216
	5. ***-дай, (-дей, -тай, -тей)*** жұрнағы жалғанып, сөздерді салыс-тыру, ұқсату мәнді туынды сын есімдер жасалса. Мыс.: *аттай, әкедей, мендей, сендей, жүздей, өлердей* т.б.	217
	6. ***-лық, -лік, -дық, -дік, -тық, -тік*** жұрнақтары арқылы мына мағынадағы туынды сын есімдер жасалса: а) зат есімдерге жалғанып, сөздердің нақтылылық мағына қасиеттерін білдіретін туынды сын есімдер жасалса. Мыс.: *орталық, қоғамдық, қалалық, азаматтық, жолдастық* т.б.	218
	ә) мезгіл атаулары мен әр қилы бұйым атауларына жалғанып, мезгіл, өлшеу мөлшерімен байланысты туынды сын есімдер жасалса. Мыс.: *айлық, жылдық, апталық, тәуліктік, көйлектік, пальтолық, қайнатым-дық, екі-үш асымдық* т.б.	219
	б) есімдіктерге жалғанып, олардан белгілі бір жаққа қатыстықты білдіретін туынды есімдер жасалса. Мыс.: *өздік, мендік, сендік, қандайлық, қаншалық* т.б.	220

	7. ***-лас, -лес, -дас, -дес, -тас, -тес*** жұрнақтары арқылы мына мағынадағы туынды сын есімдер жасалса: а) адам мінезіне байланысты зат есімге жалғанып, ондай сипаттың басқа адамдарға да қатысты екендігін білдірсе. Мыс.: *ниеттес, сырлас, тілеулес, пікірлес* т.б.	221
	ә) адамзат, жан-жануарлар тегімен байланысты ұғымдағы зат есімдермен жалғанып, тұқым-туыстас екенін білдіретін сын есім жасалса. Мыс.: *туыстас, аталас, бауырлас, қарындас, тумалас* т.б.	222
2	б) мекен ұғымдарына жалғанып, олармен қоныстастығын білдіретін мағынадағы туынды сын есімдер жасалса. Мыс.: *ауылдас, көршілес, жерлес, елдес, тумалас, ұялас, жапсарлас, іргелес* т.б.	223
	в) кейбір зат есімдерге жалғанып, олардың нақтылы лексикалық мағыналарына қарай, түрлі қарым-қатынас, мерзім-мөлшер жағынан сыбайлас келетін сындық ұғымдарды білдірсе. Мыс.: *жолдас, замандас, сабақтас, кәсіптес, қызметтес, дәмдес, өкшелес* т.б.	224
	8. ***-шаң, -шең*** жұрнақтары арқылы жасалатын туынды сын есімдер: а) киім-кешек атауларына жалғанып, адамның бойындағы киімімен байланысты сыртқы бейне-көрінісін білдіретін сын есімдер жасалса. Мыс.: *көйлекшең, етікшең, пальтошаң, шалбаршаң* т.б.	225
	ә) кейбір зат есімдерге жалғанып, адамға не затқа тән белгілі бір ерекше қасиет білдіретін туынды сын есімдер жасалса. Мыс.: *ашушаң, сөзшең, бойшаң, тершең, кіршең* т.б.	226
	9. Есімдерден туынды сын есім тудыратын өнімсіз жұрнақтар арқылы: а) ***-дар, -дер, -тар, -тер*** қосымшалары кейбір зат есімдерге жалғанып, белгілі бір іс-әрекетіне душар болғандықты білдіретін сындық ұғым жасалса. Мыс.: *қарыздар, хабардар, борыштар* т.б.	227
	ә) иран тілдерінен (парсы, тәжік, ауған) ауысқан кірме ***«-и»*** қосымшасы кейбір есімдерге жалғанып, туынды сын есім жасалса. Мыс.: *әскери, мәдени, тарихи* т.б.	228
	б) ***-паз*** жұрнағы жалғанып, сын есімге бейім жаңа сөз жасалса. Мыс.: *өнерпаз, аспаз, білімпаз, ойынпаз* т.б.	229
	в) ***-мпаз (-м+паз), -ымпаз (-ым+паз), -імпаз (-ім+паз)*** жұрнағы (және компоненттер) кейбір есім сөздерге жалғанып, туынды сын есім жасалса. Мыс.: *сезімпаз, алымпаз, білімпаз, жағымпаз, жасампаз* т.б.	230
	г) ***-қой, (-ғой, -гөй)*** жұрнағы кейбір зат есім есім сөздерге жалғанып, амал-әрекет пен мінез-құлықтардың атауы болатын туынды сын есім тудырса. Мыс.: *кәсіпқой, әзілқой, сәнқой, жәдігөй, әзілқой, сәнқой, жанжалқой* т.б.	231
	д) ***-қор*** жұрнағы зат есім есім сөздерге жалғанып, белгілі бір машықтанғанды білдіретін мағынадағы туынды сын есім жасалса. Мыс.: *жемқор, жалақор, айлақор, бейнетқор, қамқор, мансапқор, ызақор, намысқор* т.б.	232
	Етістіктерден өнімді жұрнақтар арқылы жасалатын туынды сын есім сөздер: 1. ***-қ, -к, -ық, -ік, -ақ, -ек:*** а) салт және сабақты етістікке жалғанып, түрлі сындық ұғымдағы атаулар жасайды. Мыс.: *ашық, тұнық, шірік, ілік, дөңгелек* т.б.; б) еліктеу сөздерге жалғанып, қатыстық сын есім тудырса. Мыс.: *бұлтақ, жалтақ, еңкек, жалпақ* т.б.;	317
	в) кейбір зат есімдерге жалғанып, туынды сын есім жасаса. Мыс.: *жолақ, қасырақ, ортақ, ирек* т.б.; г) етістіктен ***-уық (-уік)*** формасы арқылы белгілі бір іс-әрекетке бейімділікті білдіретін туынды сын есім жасаса. Мыс.: *жылауық, сөйлеуік, сұрауық, сыбырлауық* т.б.	

2	2. ***-ыңқы, -ңқы, -іңкі.*** Мыс.: *жатыңқы, салбыраңқы, көтеріңкі, кебіңкі, батырыңқы* т.б.	318
	3. ***-ынды, -інді, -нды, -нді.*** Мыс.: *асыранды, серпінді, жаттанды, шұбырынды, түйінді* т.б.	319
	4. ***–малы (-мелі, -балы, -белі, -палы, -пелі).*** Мыс.: *ауыспалы, көшпелі, таңдамалы, аумалы-төкпелі* т.б.	320
	5. ***-қыш, -кіш, -ғыш, -гіш***. Мыс.: *білгіш, оңғыш, тапқыш, сенгіш,* т.б.	321
	6. ***-шақ (-шек).*** Мыс.: *мақтаншақ, еріншек, қызғаншақ, ашуланшақ, жасқаншақ, тартыншақ* т.б.	322
	7. ***-ымды (-імді, -мды, -мді).*** Мыс.: *жағымды, ұғымды, үйлесімді, жарасымды, сенімді* т.б.	323
	8. ***-улы, -улі***. Мыс.: *жинаулы, ерттеулі, үюлі* т.б.	324
	9. ***–қақ, -кек, -ғақ, -гек.*** Мыс.: *асқақ, тоңғақ, жабысқақ, майысқақ, тайғақ, оңғақ, қатқақ, ұрысқақ* т.б.	325
	10. ***ма, -ме, -ба, -бе, -па, -пе.*** Мыс.: *жалдама (ақы), қызба (адам), көшпе (құм), сырма (бешпет), аспа (шам), сырма (бешпет)* т.б.	326
	Етістіктерден өнімсіз жұрнақтар арқылы жасалатын туынды сын есім сөздер: 1. ***–ыс, -іс, -с.*** Мыс.: *келіс, кетіс, ұқсас, таныс, тіркес, жалғас* т.б.	327
	2. ***–қыр, -ғыр, -кір, -гір***. Мыс.: *білгір, ұшқыр, үшкір, алғыр, өткір* т.б.	328
	3. ***–мыс (-мыш).*** Мыс.: *жасамыс, айтылмыш, тарамыс, жазылмыш, жаралмыш т.б.*	329
	4. ***–ымтал (-імтал).*** Мыс.: *ұғымтал, өсімтал, сезімтал* т.б.	330
	5. ***–қы (-кі, -ғы, -гі).*** Мыс.: *бұралқы, оралғы, жинақы, күлдіргі* т.б.	340
	6. ***–ыр (-ір, -ар, -ер, -р).*** Мыс.: *жұмыр, иір, обыр, қыңыр, тықыр, шымыр, құзар, былжыр, жалтыр, жылтыр, сылбыр, былбыр* т.б.	341
	7. ***–у.*** Мыс.: *жарау, қызу, тақау, жадау, жабырқау, түзу, таяу, бітеу, баяу, көлбеу, қату* т.б.	342
	8. ***–ған (-еген).*** Мыс.: *тебеген, қабаған, сүзеген, қашаған, береген, алаған, безеген, көреген* т.б.	343
	9. ***–ын (-ін, -н).*** Мыс.: *ортан, бүтін, еркін, ұзын, жайын* т.б.	344
	10. ***–қалақ (-келек, -ғалақ, -гелек).*** Мыс.: *ұшқалақ, сасқалақ, қозғалақ* т.б.	345
	11. ***–алақ (-елек).*** Мыс.: *еңкелек, қаңғалақ, бұлталақ, шыжалақ* т.б.	346
	12. ***–анақ.*** Мыс.: *сұғанақ, шұқанақ, шығанақ* т.б.	347
	13. ***–ғылықты (-гілікті, -қылықты, -кілікті).*** Мыс.: *жеткілікті, тұрғылықты, жергілікті* т.б.	348
	14. ***–мсақ (-мсек, -ымсақ, -імсек).*** Мыс.: *сұрамсақ, тілемсек, жарамсақ, берімсек, өлімсек* т.б.	349

Енді, осы *1-кестеде* қарастырылған сын есімнің лексика-морфологиялық құрылымына ақпараттық белгі-кодты сәйкестендіру бағдарламасы негізінде, кейбір негізгі және туынды сын есімдерге қатысты сөздерге де шартты белгі-код сәйкестендіру мүмкіндігі бар.

Әдебиет

1. *Ысқақов А.* Қазіргі қазақ тілі. Алматы: «Ана тілі», 1991. 384 б.
2. *Головин Б.Н.* Язык и статистика. М., 1971. 191 с.
3. *Жұбанов А.Қ.* Қолданбалы тіл білімінің мәселелері. Вопросы

прикладного языкознания. – Алматы: «Арыс» баспасы, 2008. – 640 бет.
4. *Жұбанов А.Қ.* Қазақ тілі корпусындағы зат есім сөздерге лексико-морфологиялық деңгейдегі белгіленім қою бағдарламасы // Қазақ тілінің тарихы мен түркітану мәселелері /Жауапты редактор ф.ғ.д., профессор М.Малбақов. –Алматы: А.Байтұрсынұлы атындағы Тіл білімі институты, 2011, – 228-233 бб.
5. *Жұбанов А.Қ.* Қазақ тілінің аннотацияланған мәтіндер корпусындағы етістік сөздерге лексика-морфологиялық белгі-код қоюдың (белгіленім) алғышарттары. «Тілтаным» журналы. 2012. № 1. 18-25-бб.

СЕМАНТИЧЕСКИЕ КЛАССЫ СУЩЕСТВИТЕЛЬНЫХ, ИСПОЛЬЗОВАННЫХ В МОДЕЛИ ПОРОЖДЕНИЯ КАЗАХСКОГО ТЕКСТА

ҚР БҒМ Оңтүстік Қазақстан Мемлекеттік
педагогикалық институты «Елбасы жолдауы
және тіл біліміндегі жаңа ізденістер» атты
халықаралық-тәжірибелік конференция
материалдары. шымкент – 2012, - б. 112-118.

В соотношении «язык – реальная действительность» язык служит не только для обозначения предметов, фактов и явлений реальной действительности. Он, в то же время, является носителем обобщенного отражения действительности, органически соединяясь с соответствующими понятиями или отдельными их признаками, лежащими в основе семантических ценностей языковых единиц.

Общеизвестно, что лексические средства, наряду со звуковыми и грамматическими, составляют основу любого языка, в частности, и казахского. Слово, называя предметы, явления реального мира в их обобщенном виде, выступает в языке основным средством номинации. Являясь структурной ячейкой строя языка, оно имеет определенную морфологическую оформленность и грамматическую значимость в языке. Слово охватывает фактически весь объем языковых функций: номинативную (обозначения), сигнификативную (обобщения), коммуникативную (общения) и прагматическую (экспрессивного выражения и эмоционального воздействия). Словарные знаки слов таких классов, как существительные, глаголы и прилагательные, обладают двумя основными семиологическими функциями – наименования и описания. Совместимость их в пределах слова порождает назывные высказывания, предикативные и лексические синтагмы. Они являются тем основным звеном в словарном составе языка, той постоянной значимой основой, которая обеспечивает “переход” от языка к мысли, от системы к речи, от слов к синтагматическим единицам, большим, чем слова [1, С.82-100].

При семантической классификации лексики казахского языка (выделении в ее составе определенных лексико-семантических классов и подклассов) имеются трудности и специфика. Это одна из труднейших задач семантики, которая не имеет единого решения и в других языках мира. Количество таких классов и подклассов и их состав зависят от целей исследования, лингвистической школы и ряда других факторов.

Работ, посвященных иссследованию казахского языка в области семантики, относительно мало. Здесь стоит отметить исследования М.Оразова и его монографию «Қазақ тілінің семантикасы» и другие его работы [2÷5], которые затрагивают многие вопросы семантики казахского языка. В теоретических и прикладных аспектах казахстанскими учеными проводятся различные исследования, затрагивающие семантические стороны языковых единиц казахского и русского языков. Большую научную ценность имеют интерлингвальные исследования Э.Д.Сулейменовой по установлению семантических параметров слова в аспекте контрастивного анализа [6].

Особого внимания заслуживают научные изыскания З.К.Ахметжановой, посвященные функционально-семантическим полям русского и казахского языков. Применение функционального подхода, основных понятий теории функциональной грамматики к сопоставительному анализу типологически контрастных языков смело можно назвать началом нового этапа в развитии сопоставительного языкознания, так как здесь открывается возможность изучения системы языка в действии, в живом употреблении, т.е. сопоставительно-типологического исследования нескольких языков в функциональном аспекте. Кроме того, исследования З. К. Ахметжановой позволяют говорить о текстоорганизующей роли посессивности казахского языка и об отсутствии таковой функции у русского языка [7].

В самом общем виде можно выделить три принципа семантической классификации лексики казахского языка [8]: 1) денотативно-десигнативный; 2) парадигматический; 3) синтагматический.

Денотативно-десигнативный принцип учитывает онтологическое расчленение действительности. При этом слова объединяются в группы в соответствии с тем, что они обозначают в реальной действительности («названия животных», «названия растений», наименования определенных классов предметов и т.п.). При *парадигматической классификации* слова объединяются в группы на основе тождества определенных семантических компонентов, выделенных в значении слов («модальные глаголы», «каузативные глаголы» и т.п.). По синтагматическому принципу слова объединяются в лексико-семантические группы с учетом количества и качества валентностей классифицируемых слов [5, с.189-192].

Исходя из целей нашего исследования, в основу семантической классификации положен *денотативно-десигнативный принцип*, хотя в некоторых случаях он будет дополняться двумя другими, перечисленными выше.

Семантическая классификация имен существительных казахского языка. Как известно, основным критерием разграничения имен существительных казахского языка на отдельные группы, как и в других языках, является, прежде всего, анализ их прямого номинативного значения, детерминированного типом денотата и сигнификата. При этом, “слово” можно рассматривать двояко, т.е. как единицу первичного означивания в системе номинативных средств языка и как единицу, семантически реализованную в речи. Нами различаются две модификации «денотатов»: денотат как объект обозначения единичного предмета или класса предметов, а также денотат как предмет референции знака в речи.

А. А. Уфимцева, в книге «Типы словесных знаков» [1, с.102], трактует сигнификат в шести понятиях как компонент знакового значения слов, что, как нам кажется, справедливо и для казахского языка:

1. Содержание полного конкретного понятия, противопоставленное его объему (денотату), как это имеет место у имен классов конкретных или единичных предметов, явлений. Например: *қала (город), жел (ветер), кітап (книга)* и т.п.

2. Абстрактное понятие признака, свойства, действия, отношения и т.п., в высшей степени обобщенное и непротивопоставленное предметному представлению. Например, *данышпандық (гениальность), ер жету (зрелость), надандық (тупость), қараңғылық (темнота), суықтық (холодность), жан тыныштық (спокойствие)* и т.п.

3. Абсолютные, в высшей степени обобщенные понятия, выражающие мыслительные категории. Например, *біртуар (феномен), белгі (знак), нышан (признак), қасиет (свойство), сан (количество)*.

4. Абстрактное понятие, обозначающее категории реального мира и выступающее родовым по отношению к другим. Например, *сезім (чувство), оқиға (событие), қозғалыс (движение), єрекет (действие), сұйықтық (жидкость)* и т.п.

5. Единичное понятие, представляющее собой «мыслительный конструкт» ирреального предмета. Например, *кентавр, леший, русалка* и т.п. В казахском языке такие, как *диуана, пері, жын-шайтан* и т.п.

6. Метаязыковые понятия, термины различных областей знаний (науки, техники, искусства и т.п.).

Из перечисленного вытекает, что имена с денотативным и денотативно-сигнификативным значением относятся к разряду имен, называемых конкретной лексикой, а слова с сигнификативной и сигнификативно-денотативной основой значения составляют абстрактную лексику любого языка. Необходимо заметить, что определяющим для семантики именных лексем является обязательное наличие в их знаковом значении денотата и сигнификата. Если словесный знак служит специализированным названием (термином) в какой-либо узкой предметной области, то доминирует денотат над сигнификатом. В случае, когда словесный знак служит названием абстрактного понятия, то, наоборот, или сигнификат доминирует над

денотатом, или словесный знак имеет чисто сигнификативное значение [1, с.103].

Именами существительными в казахском языке, так же, как и во многих тюрских языках, являются различные наименования предметов и понятий в широком смысле слова. Это, например, различные имена собственные, такие, как имена людей, наименования городов, рек, озер и т. д.; нарицательные имена – названия человека: *кісі (человек), шеше (мать), єйел (женщина);* названия животных; растений; названия различных веществ и предметов быта; научные, технические и другие термины; наименования разных обобщенных отвлеченных понятий: *ел (народ), ру (род), тағдыр (судьба)* и т.д.; наименования действий, явлений, состояний: *жүріс (ходьба), хабар (известие), ұйқы (сон)* и т.д.; названия результатов действия: *жиналыс (собрание; ср. жинау – собираться), кеспе (лапша; ср. кесу – резать);* названия орудий (инструментов) действия: *бұрғы (сверло; ср. бұрау – винтить, завинчивать), сыпырғыш (метла, веник), ара (пила)* и т.п.

В отличие от европейских языков, в казахском языке имена существительные не различаются грамматически по признаку одушевленности или неодушевленности. В семантическом отношении только наименования людей различаются как одушевленные, отвечая на вопрос *кім? (кто?),* а все остальные (живые и не живые) отвечают на вопрос *не? (что?).* В случае необходимости различения лиц мужского и женского пола животного мира используются понятиями пола по значению самого слова.

Корневые имена существительные в казахском языке в основном совпадают с корнем и, обладая полнотой значения, соответствуют самостоятельному слову. Последнее намного облегчает процесс автоматического расспознавания их в тексте. Термин “производная основа” может быть применен в казахском языке, главным образом, в отношении глаголов и отглагольно-именных образований. Таким образом, имена в казахском языке являются самостоятельной лексико-грамматической категорией и в предложении могут употребляться без дополнительных аффиксов [9].

Как пишет М. Оразов, среди слов казахских существительных самыми частыми по сложности их значений являются нарицательные имена. Эти имена существительные по отношению к значению имеют в языке изменчивый характер, т.е. иногда у них превалирует денотативное значение, а иногда сигнификативное или приобретается как бы суммарное значение – и денотативное, и сигнификативное. Примером может служить использование слов, которые применяются в метафорическом и метомических значениях. Безусловно, когда слова употребляются в переносном значении, то изменяется его и денотативное, и сигнификативное значение. Для убедительности изложенного можно сравнить слово «нар» в двух следующих предложениях из стихов: *қырық балам бар еді, қырқы да бірдей нар еді* (Батырлар жыры). *Айырдан туған жампоз бар – нарға жүгін салғызбас* (Махамбет) [5, с.98]. В первом предложении слово «нар» применено в переносном значении, т.е. как сильный, «боевой», а во втором – в прямом

значении, как одногорбный верблюд. Их денотативные и сигнификативные значения несравнимы.

В количественном отношении тематические группы имен существительных в любом языке, в том числе, и в казахском, довольно велики. М. Оразов в книге «Қазақ тілінің семантикасы» приводит в виде примера 8 разновидностей таких тематических групп [5, с.191]: 1) родственные термины *(туысқандық атаулар);* 2) человек и названия его части тела *(дене мүше атаулары);* 3) названия домашних животных *(төрт түлік мал атаулары);* 4) названия одежды, обуви человека *(киім-кешек атаулары);* 5) название птиц *(құс атаулары);* 6) название диких животных (жабайы аң атаулары); 7) название обиходных предметов *(ыдыс-аяқ атаулары);* 8) названия болезни человека *(адам баласына тән ауру атаулары).*

С учетом целей нашего исследования, анализиуемого материала, а также отдельных идей, нами выделены следующие семантические классы существительных, которые приведены в ***табл. 1***. В нашем исследовании в основу семантической классификации казахских существительных положено деление реальной действительности на определенные зоны, сегменты [10]. Как видно из этой *табл.1*, денотативно-десигнативный принцип использован здесь при выделении наиболее крупных семантических классов существительных. Более мелкие лексико-семантические группы, например «части тела человека», «воинские звания», «дни недели» и др., связаны парадигматическими отношениями.

Таблица 1. **Семантические классы существительных, использованных в модели порождения казахского текста**

Сегменты реальной действи-тельности	**Наименование семантического класса**	**Код класса**
1	**2**	**3**
1. Наука	Общефилософские понятия	ЗТ101
	Общенаучные понятия	ЗТ102
	Общемедицинские понятия	ЗТ103
	Основные понятия психологии	ЗТ104
	Основные понятия нейропсихологии	ЗТ105
	Основные понятия логики	ЗТ106
	Основные понятия геометрии	ЗТ107
	Основные географические понятия	ЗТ108
	Основные понятия языкознания	ЗТ109
	Основные понятия литературоведения	ЗТ110
	Основные понятия изобразительного искусства	ЗТ111
	Основные понятия музыкального искусства	ЗТ112
	Основные юридические и уголовные понятия	ЗТ113
	Основные понятия, связанные с театром	ЗТ114
	Понятия, связанные с авиацией	ЗТ115
	Понятия, связанные со строительством	ЗТ116

	Понятия, связананные с устройством отдельного механического объекта	ЗТ117
	Информационные понятия	ЗТ118
	Собирательные понятия	ЗТ119
	Основные понятия электроники и электротехники	ЗТ120
	Административные понятия	ЗТ121
	Имена организаций, газет, журналов и т.д.	ЗТ122
	Абстрактные имена – мыслимые образы	ЗТ123
	Оценка	ЗТ124
2. Природа	Предметы природы	ЗТ201
	Явления природы	ЗТ202
	Предметы природы – артефакты	ЗТ203
	Состояния природы	ЗТ204
	Понятия, описывающие структуру предметов и части предметов	ЗТ205
	Вещества, металлы	ЗТ206
	Действия предметов природы	ЗТ207
3. Человек и его мир	Человек и его части тела	ЗТ301
	Части лица и головы человека	ЗТ302
	Состояния и ощущения, специфичные для человека	ЗТ303
	Действия, спецецифичные для человека	ЗТ304
3. Человек и его мир (продол-жение)	Понятия, используемые при характеристике человека	ЗТ305
	Предметы одежды и обуви человека и понятия, с этим связанные	ЗТ306
	Предметы-артефакты, используемые человеком в целях гигиены, для украшения и т.п.	ЗТ307
	Объекты, исполь-ые для передвижения человека и понятия, с этим связанные	ЗТ308
	Предметы питания человека	ЗТ309
	Понятия, связанные с жилищем человека	ЗТ310
	Понятия, связанные со смертью и захоронением	ЗТ311
	Имена, отчества и фамилии людей	ЗТ312
	Национальность человека	ЗТ313
	Предметы-сосуды, которыми пользуется человек, и связанные с этим понятия	ЗТ314
	Предметы, используемые на отдыхе, охоте, рыбалке, и понятия, с этим связанные	ЗТ315
	Наименование гражданских специальностей, должностей, степеней, званий	ЗТ316
	Предметы-артефакты, используемые при учении, труде и т.п.	ЗТ317
	Понятия, связанные с взаимоотношением людей	ЗТ318
	Понятия, связанные с родственными отношениями между людьми	ЗТ319
	Награды человека	ЗТ320
4. Живот-ный мир	Крупные домашние животные и части их тела	ЗТ401
	Мелкие домашние животные и части их тела	ЗТ402
	Крупные дикие животные и части их тела	ЗТ403
	Насекомые и части их тела	ЗТ404
	Птицы и части их тела	ЗТ405
	Действия, спецефичные для животных	ЗТ406
	Предметы питания животных, птиц и насекомых	ЗТ407
5. Растения	Растения и их части	ЗТ501

	Деревья и их части	ЗТ502
6. Время	Временные понятия	ЗТ601
	Дни недели	ЗТ602
7. Религия	Понятия, связанные с религией	ЗТ701
8. Война и понятия, с ней связанные	Предметы боевой техники, оружия и их части	ЗТ801
	Наименование военных специональностей	ЗТ802
	Воинские звания	ЗТ803
	Понятия, связанные с оказанием помощи во время войны	ЗТ804
	Названия воинских подразделений	ЗТ805
	Понятия, связанные с войной	ЗТ806
9. Учеба и труд	Организация учебы, труда	ЗТ901
	Понятия, связанные с обучением и трудом	ЗТ902
10. Меры	Меры длины, веса, скорости и т.п.	ЗТ911
11. Геогра-фия	Имена населенных пунктов	ЗТ921
	Имена стран, материка, частей света, островов, гор и др.	ЗТ922
	Имена рек, морей, озер, проливов	ЗТ923

В последней графе ***табл. 1*** приведены принятые в модели порождения коды семантических подклассов существительных. ЗТ – сокращение от *зат есім* (существительное).

Литература

1. *Уфимцева А. А.* Типы словесных знаков. - М.: Наука, 1974. - 206 с.

2. *Оразов М.* Семантика казахского глагола. (Опыт семантической классификации.): Дис. докт. - Алма-Ата, 1983. - 415 с.

3. *Оразов М.* Қазіргі қазақ тіліндегі қалып етістіктері. - Алматы, 1980.

4. *Оразов М.* Сөз теориясы. - Ташкент, 1985.

5. *Оразов М.* Қазақ тілінің семантикасы. - Алматы: "Рауан", 1991. -215 б.

6. *Сулейменова Э. Д.* Семантические параметры слова в интерлингвальном исследовании (теоретические и прикладные аспекты контрастивного анализа). АДД., М.: 1992.

7. *Ахметжанова З. К.* Функционально-семантические поля русского и казахского языков. - Алма-Ата: Наука, 1989. - 106 с.

8. *Васильев Л. М.* Речевой смысл предложения как результат взаимодействия языкового и неязыкового знания. - В сб.: Предложение и текст: семантика, прагматика и синтаксис. - Л.: изд. Ленингр. унив. 1988. - с. 5-10.

9. *Ысқақов А. Ы.* Қазіргі қазақ тілі. Морфология. - Алматы: Ана тілі, 1991. - 382 б.

10. *Зубов А. В.* Вероятностно-алгоритмическая модель порождения текста (семантико-синтаксический аспект). - Минск.: 1985. - 399 с. (докторская диссертация, филолог. наук).

ТҰРАҚТЫ СӨЗ ТІРКЕСТЕРІН КОДТАУ ЖӘНЕ МОДЕЛЬДЕУ МӘСЕЛЕСІ

ҚР БҒМ М.Сапарбаев атындағы Оңтүстік Қазақстан
гуманитарлық институты Академик М.Б.Сапарбаевтың
70 жылдығына арналған «Жоғары оқу орнында жастарға
заманауи білім мен тәрбие берудің өзекті мәселелері»
атты халықаралық ғылыми-тәжірибелік конференцияның
материалдары. Шымкент, 2012. 261-263-бб.
(Т. Садықов және М.Ермекбаевпен бірге)

Лингво-математикалық үлгілеу мәселесі әлі күнге дейін қазіргі заманғы инженерлік және квантитативтік лингвистиканың маңызды мәселелерінің бірі ретінде қарастырылып келеді. Бұл мәселе, бір жағынан, лингвистикалық нысандардың континуалды аморфтылығының қарама-қарсы қойылуымен және олардың математикалық ақпараттарының нақтылы дискретті үлгілеуімен байланысты. Екінші жағынан, табиғи тілдердегі таңбалардың мағыналық байлығының арасындағы сәйкессіздікпен және математикалық таңбалардың салыстырмалы семантикалық жұтаңдығымен да байланысты.

Қазіргі таңда аталған мәселені жан-жақты зерттеу және шешімін іздеу үшін флективті және флективті-түбір үндіеуропа тілдері лингвистикалық нысан ретінде қолданып келеді. Бұл тілдер (флективті және флективті-түбір) жоғарыда сөз етілген тілдерден айқын және тұрақты құрылымы жағынан ерекшеленеді. Бұның себебі, біріншіден, бұл тілдерді қазіргі заманауи кибернетиканың көмегімен тілдік материалдар ретінде қарастыруға болады. Екіншіден, жаңа геосаяси жағдайда бұрынғы Кеңес Одағының территориясында құрылған жас мемлекеттердің тілдік құрылысы ұлттық тілдердің қолданылу және қызмет аясының кеңеюіне назар аударады. Бұл міндетті есептеу техникасының көмегінсіз орындау мүмкін болмайды.

Үшіншіден, жас мемлекеттердің жекелік сыртқы экономикалық және мәдени саясатының тағдыры қостілдік машиналық аударма мен референттеу жүйелерін қолданумен тығыз байланысты. Сондықтан, орыс және батыс еуропа тілдерінде мұндай жүйені құрудың отыз жылдық тәжірибесіне сүйене отырып, ғалымдар алдына қостілді аударма және референттеу жүйесін құрастыру міндетін қойып отыр.

Орыс тіліне қатысты инженерлік лингвистиканың отыз жылдық тәжірибесі қостілді машиналық аударма, автоматты индекстеу, аннотациялау, референттеудің қазіргі заманауи жүйелері квантитативті лингвистиканың әдіс-тәсілдерін қолданбай, сондай-ақ осы тілдердің ақпараттық қорын құрмай жүзеге асуы мүмкін емес. Сол себепті, біз қазақ тілі (нақтырақ айтсақ, лексикалық аспектісі) үшін үлгілеу тәсілдерінің жүйесін құруға мәжбүр болып отырмыз. Себебі, тікелей қарапайым бақылау арқылы қазақ тілін басқа қыпшақ тілдерімен салыстыруға мүмкіндік беретін ішкі қасиеттерін анықтау мүмкін емес. Мұндай үлгілеу – лексиканың статистикалық сипаты. Бәрімізге белгілі, тілде бұрын жоқ болған жаңа нәрсе болмайды («nihil est in lingua guod non prius

fuerit in oration» – Бенефист). Сондықтан, тілді қазіргі заманға лайықты зерттеу және қарастыруды осы жүйелер арқылы пайда болған мәтіндердің көмегімен бастау керек. Белгілі бір тілдің мәтіндерін жан-жақты зерттеудің қолданыстағы тәсілінің бірі – оның құрылымы, содан кейін квантитативті және ақпараттық үлгілерінің талдауын жүргізу. Философиялық тұрғыда үлгілеудің мағынасы – табиғатты танудың тиімді құралы. Тілдік шығармалардың әлемді үлгілеуге қатыстылығын, үлгілердің жалпы қасиеттерін талдау мен олардың тілдік шығармалар талабына сай келуін анықтау негізінде көз жеткізуге болады. Ал «Табиғи тіл – бізді қоршаған шындықтың үлгісі» деген тұжырымды қазіргі кездегі тілшілердің көпшілігі қолдайды. Дегенмен, мұндай тұжырыммен тек тілдік шығармалар (немесе мәтін) жайлы сөз болған жағдайда ғана келісуге болады. Негізі, «модель» термині латынның «modulus» деген сөзінен шыққан, мағынасы – «өлшем». Модель сан алуан болады, тіпті бір ғылым саласының ішінде де «модель» терминін әртүрлі мағынада қолданылады. Модель деп бір заттың үлгісін айтады немесе сол заттың өзі сияқты ұқсас түрлі (кішірейтілген, үлкейтілген немесе табиғи өлшемінде, түрінде, мысалы, ұшақтың моделі немесе алмастың кристалдық торының моделі) жасандыны айтады Ал үлгілеу дегеніміз бір нысанның екінші нысанның орнын басып, сол арқылы модель – нысанның көмегімен түпнұсқа нысанының аса маңызды қасиеттері туралы ақпарат алу. Яғни модель – түпнұсқаның бейнесі. Егер түпнұсқа мен модельдің арасындағы сәйкестіктер сай келіп жатса, модельді дәлелді деп санауға болады. Табиғи тілде референт пен сигналдың арасындағы байланысты адам баласы сыртқы әлемдегі танымдық іс-әрекеттерінің барысында белгілейді. Адамдар танымдық іс-әрекетте қоршаған ортаны таниды, жаңа ақпаратты меңгеруге талпынады. Таным үдерісі – күрделі процесс. Ғалымдар «Таным – адам санасын дамытудың негізі» [1; 2] және «Таным – бұл қоршаған болмыстың, оның элементтерінің белсенді ақыл-ой және көңіл күй іс-әрекеттері мен олардың нәтижесі, яғни білім, жалпыланған теория, заңдар мен ғылыми ұйымдардың санада бейнеленуі» [3] деген пікірде. Олай болса адам санасында осы шартқа байланысты бекітілуінің арқасында ғана сыртқы әлемнің бір объектісі, мәселен, айтылған немесе жазылған сөз адам баласы үшін этикет немесе сыртқы әлемнің басқа бір объектісінің сипаттамасы болуы мүмкін. Әрбір таңба – екі жақты психикалық мән. Ол сигналдың психикалық көрінісі болып табылатын анықтаушы (біздің жағдайда, ешкі) және адам санасында сыртқы әлемнің шынайы объектісінің орнын алмастыратын анықталушыдан (біздің жағдайда, ешкі) тұрады. Біздің мақсатымыз тұрақты сөз тіркестерінің сөзжасамдық кодтарын, үлгілерін құру. Мұнда А. Жұбанов, Б. Әбікеева, Н. Нұрлыбаевтардың еңбектері негізге алынды.

Үлгілеуде кодтар тікелей мәтінге қатысты үдерістерді іске асырады, алдын ала өңделіп, жорамал ретінде алынған сөзжасамдар сөзтұлға кластарының таңбаларына ауыстырылады, яғни қайта кодталады. Қайта кодтауда арнайы таңбалар қолданылады, ал егер осы таңбаларды қазақ тіліне ыңғайлайтын болсақ, онда белгілеу таңбаларының келесі тізбегі шығады:

1. N_a – атау септігіндегі зат есім мен жіктеу есімдігі;
2. N_i , $N_ж$, $N_б$, $N_ш$, $N_к$, $N_т$ – басқа қалған септіктер;

3. A_a , A_i , $A_ж$, $A_б$, $A_ш$, $A_к$, $A_т$ – атау, ілік, жатыс, барыс және т.б. септіктердегі сын есім және оның эквиваленттері;

4. ***V*** **–** етістік, жіктік жалғау;

5. ***P*** – септеуліктер;

6. $P_ж$, $P_б$, $P_т$ – ілік септігінен соң, 3-жақтағы тәуелдік жалғауынан кейінгі жатыс, барыс, табыс септіктер;

7. С – шылаулар және т.б.

Математикалық шартты белгілер ағылшын, орыс тіліндегі ғылыми әдебиеттерінде көптеп кездеседі. Осы берілген ереже бойынша қазақ тілі **тұрақты сөз тіркестерінің сөзжасамын** кодтауға болады. Екі сөзден құралған тұрақты сөз тіркестері болса, мысалы,

1. ***N+N*** **(есім сөз+есім сөз) формасындағы тұрақты сөз тіркестері:** темір жол;

2. **Анықтауыштық қатынастағы тұрақты сөз тіркестері** – ***A+N*** **(анықтаушы мағынадағы есім сөз+есім сөз) моделі:** жомарт жан, аялы алақан;

3. ***N + V*** **(есім сөз + етістік)**: Жүрегі ұшты.

Үш немесе одан да көп сөздерден құралған тұрақты сөз тіркестеріне келесі белгілер қолданылады. Мысалы: ***N + V*** (есімше) + ***S*** **формасындағы тұрақты сөз тіркестері:** Ит өлген жерде $N_a V N_ж$, ***Көзді ашып жұмғанша*** – $N_m V_{көс} V_{ес}$.

Бұл құрылған модельдерді «тар» мағыналы модельдерге жатқызуға болады. Себебі мұнда тұрақты сөз тіркестерінің грамматикалық минимум мағынасын тек бөлек-бөлек алынған сөз таптары ғана ашып тұр. Түркі тілдерінде, соның ішінде қазақ тілінде сөз таптары лексика-грамматикалық категориялар болып табылады. Барлық сөз таптарын анықтауға арналған негізгі өлшемдер қызметін атқаратын сөз жасаушы қосымшалар бір сөз табын екінші сөз табынан формалды түрде жіктеп тұратын мәнді жәпе айқын белгілер болып табылады. Ал тұрақты сөз тіркестерінің лексикалық мағынасын кодтау мәселесі келесі еңбектерде жан-жақты суреттеледі.

Әдебиеттер

1. *Беляев Б.В.* О взаимоотношении мышления, языка и речи // Вопросы психологии, 1958.

2. *Леонтьев А.А.* Язык, речь и речевая деятельность. - М.: Наука, 1969. -361 с.

3. *Беляев Б.В.* Психологические закономерности владения иноязычными словами. *В кн.:* Психология и методы обучения второму языку. –М., 1987. -375 с.

ҚАЗАҚ ТІЛІНІҢ АННОТАЦИЯЛАНҒАН МӘТІНДЕР КОРПУСЫНДАҒЫ ЕТІСТІК СӨЗДЕРГЕ ЛЕКСИКА-МОРФОЛОГИЯЛЫҚ БЕЛГІ-КОД ҚОЮДЫҢ (БЕЛГІЛЕНІМ) АЛҒЫШАРТТАРЫ

А.Байтұрсынұлы атындағы Тіл білімі институтының
«Тілтаным» журналы. 2012. № 1. 18-25-бб.

Қазіргі қазақ тіл білімі заман өзгерісіне сай, жаңа технологияны – компьютерлер мүмкіндіктерін ұлттық тілді зерттеу барысында кең түрде қолдануды қажет етуде. Осыған байланысты А.Байтұрсынұлы атындағы Тіл білімі институтында «Компьютерлік лингвистика» мен «Корпустық лингвистика» салаларына тікелей қатысты бар «Қазақ әдеби тілінің аннотацияланған ұлттық корпусы» атты тақырыбы бойынша 2012-2014 жылдар аралығына жоспарланған ғылыми-зерттеу жұмысы жүргізіле бастады. Ал бұл зерттеу жұмысының негізгі мақсаты – қазақ тілінің мәдени құндылығы болып саналатын толық мәтіндер корпустарының мәтіндеріне қажетті деген грамматикалық белгі-кодтар енгізіп, оның дербес түрдегі «Тіл – қазына» атты компьютерлік базасын құру [1-3].

Осыған байланысты, кез келген қазақ тіліндегі мәтіндер корпусы аннотациямен (шартты түрдегі грамматикалық белгілермен) жабдықталады, яғни кез келген аннотацияланған корпус, оның лингвистикалық сипатының көрінісін айқындай алады. Әдетте ондай сипаттама корпус құрамындағы бірліктерге қатысты мүмкін болатын тілдік талдау үлгісін қамтуы керек. Аннотацияланған корпус, оны пайдаланудың қарапайымдылығы мен көпфункционалдық қасиеттеріне ие болуымен де аса құнды. Көпшілік жағдайда, корпустарды аннотациялау әрекеті автоматты және «қолмен» жүргізу түрлері қатар жүргізіледі. Ал автоматтанған аннотациялау әрекетін қанағаттандырарлық деп санау үшін қателік шегі үш пайыздан аспау керектігін ескеру қажет болады.

Мақалада етістік сөздерге қатысты лексика-морфологиялық белгі-код қою (белгіленім) алғышарттары қарастырылады.

Етістік сөздер құрамының басқа сөз таптарына қарағанда күрделілігі оның аса өрісті лексика-семантикалық сипатымен, өте бай лексика-грамматикалық формаларымен және кең синтаксистік қызметімен тығыз байланысты. Етістіктің лексика-грамматикалық тұлғаларының бай болу себебі – олар әрекеттің болу мезгілін, жүзеге асу кезеңін, өту сипатын, яғни әрекеттің бағыты, қарқыны, тынуы тәрізді жайларды түгел қамтиды. Осымен қатар басқа сөз табына қатысты сөздерден етістік тудыратын синтетикалық, аналитикалық тәсілдер жүйелерін қоса есептесек, етістіктің тұлғалану байлығына жететін сөз табы жоқ деуге де болады. Осы айтылғандар оның өзіне ғана тән лексика-семантикалық, лексика-грамматикалық және грамматикалық категорияларынан айқын көрініп жатады. Етістіктің осындай әрқилы категорияларының мағыналары да, синтаксистік қызметтері де өзге сөз таптарына тән сөздермен қарым-қатынасқа түскенде айқындала түседі.

Етістіктің лексикалық жүйесіндегі әрбір сөздің өзіне ғана тән лексикалық мағыналары бөлек болғанымен, семантикалық ерекшеліктеріне қарай бір сөз тобына телінеді. Етістіктің грамматикалық тұлғалары мен қызметтерін дұрыс айқындау үшін, оларды түбір формалардың, туынды синтетикалық және аналитикалық формалардың семантикалық құрылымына

орай жіктеп, топтап барып сыр-сипатын ашу, олардың тұлғалық және мазмұндық арақатынастарын айқындауға әкеліп соғады. Себебі грамматикалық семантика сөздің құрамына көп байланысты болып келеді.

Сол себептен етістік формаларының морфологиялық құрылымының сыр-сипатын ашу контекстік тұрғыда немесе мәтіндік лингвистика саласының зерттеу аясына жақынырақ келеді. Ал бұл тұрғыдан зерттеуді ұйымдастыру дегеніміз статистикалық лингвистика әдістерін молынан қолдану деген сөз. Туынды етістіктердің семантикалық құрылымының аса күрделі және түрлену сипатының бай болу себептерінің бірі – олардың сирек не жиі қолдануында деп түсіну қажет [4]. Сондықтан туынды етістіктердің морфологиялық құрылымын жеке бір жүйе ретінде қарастырып, әрбір өзіне тән сипатына қарай топтап, олардың ықтималды-статистикалық заңдылықтарын ашу үшін арнайы белгі-код жүйесін енгізіп зерттеу қажет деп табамыз. Бұдан бұрынғы мақаламызда [5] зат есім сөздерге қатысты айтылғандай, қазақ тілінің тілдік бірліктерін, сөз таптарын формальды белгілері арқылы танып білу және соның нәтижесінде автоматты түрде жазба не сөйлеу мәтінінен оларды бөліп алу әлі де іске аспай отыр. «Қолдан» енгізген шартты «белгі-код» негізінде етістіктің морфологиялық құрылымына статистикалық зерттеулер жүргізу көптеген мәселелердің басын ашатыны белгілі. Сол мәселелердің ішіне мәтіндік бірліктерді автоматты түрде ажырата білу мен олады бөліктеп алу мүмкіндігінің де қатынасы бар. Сонымен бірге етістіктің лексика-грамматикалық құрылымының дәстүрлі әдіспен зерттеуде байқала бермейтін тұстарының да басы ашылуы (анықталуы) мүмкін. Осы тұрғыда қарастыра отырып, біз төменде қазіргі қазақ тіліндегі етістіктердің (түбір және туынды) сөздерін іштей бір-біріне мағына жағынан жақындықтарына, өзара функция жағынан орайластықтарына және лексика-морфологиялық құрылымына қарай топ-топқа бөліп, олардың сипаты жайлы ақпараттарды арнайы ойластырылған шартты белгі-кодтарға сәйкестендіру бағдарламасын қарастырамыз.

Мысалы, мағыналық тұрғыдан алып, етістік сөздерді іштей мына топтарға бөлуге болады [6, 223-226-бб.]: **амал-әрекет, қимыл-қозғалыс, қалып-сапа, ойлау-сөйлеу, өсу-өну етістік-тері** және т.б. *1-кестеде* көрсетілгендей жалғасын табады.

Әдетте, туынды сөздердің бір жүйе бойынша қалыптасқан белгілі морфологиялық құрылымы болады. Сол құрылымның бірінші компоненті дербес мағыналы сөз болады да, ал екінші компоненті – бірінші компонентті белгілі бір сөз табына айналдыратын жұрнақ болады *(тіс+те, кел+тір, сабын+да)*. Туынды етістіктер, жасалатын негіздеріне қарай (бірінші компоненті), есімдерден және етістіктерден жасалған етістіктер деген екі салаға бөлінеді. Олар есім негізді етістіктер, етістік негізді етістіктер деп аталып жүр. Біз осылардың тек алғашқысын ғана, яғни есім негізді етістіктердің жасалу жолдарын жекелеп 16 топқа бөліп қарастырып, оларды шартты белгі-код арқылы таңбалау үлгісін ***1-кесте*** арқылы ұсындық.

Зат есім сөздерге лексика-морфологиялық ақпаратты сәйкестендіру бағдарламасы тәріздес, ***1-кестедегі*** бағдарлама да үш бағанадан тұрады:

1) белгі-код орны;

2) есім негізді туынды етістік сөздердің сипаты;

3) белгі-код таңбасы.

Бірінші бағананың белгі-код орны «1» (немесе №1-ші топтама), ал екінші бағана бойынша оның сипаттамасы – «етістік сөз» деп аталады егер мәтін бойынан зерттелетін кезекті сөз – *етістікке* жататын болса, онда үшінші бағанадан сол етістік сөздің *белгі-коды* – «ЕТ» екенін білеміз. Екінші орынға немесе екінші топтаманың екінші бағанасында етістіктердің өздерін іштей мағыналық және өзара функционалдық жағынан жақындастыратын 11 топтың әрбіреуіне сипаттама берілген. Ал үшінші бағанада осы сипаттамаларға сәйкестендірілген белгі-код таңбасымен танысуға болады. Мысалы, белгі-кодтың «2-ші» орнының (2-ші топтамасының) үшінші бөлігі *қалып-сапа етістіктері* деп аталады да, ал оған сәйкес келетін үшінші бағанада сол сипаттаманың үш орынды белгі-коды – «102» берілген. Сол сияқты, туынды етістіктердің жұрнақтар арқылы есім сөздерден жасалу жолынан мысал келтірейік. Туынды етістіктердің ***-ла (-ле, -да, -де, -та, -те)*** жұрнағы арқылы жасалу ішкі 5 бөлікпен топталған атаулар сипатымен берілген:

1) дене мүшелері (*аяқта, жұдырықта, өкшеле*);

2) еңбек құралдары (*арала, балтала, арқанда);*

3) іс-әрекетті жүзеге асыруға объект болатын зат атаулары (*жүнде, майла, тұзда, алтында);*

4) мекен я орын, өлшеу, дыбыс т.т. (*мекенде, өрле; аршында, метрле; мыңқылда, шыңқылда);*

5) сын есім, сан есім, үстеу, еліктеуіш т.т. сөздерге жалғанған жағдайлар (*ақта, қарала, екеуле, аһла, үһле)*.

Осы 5 түрлі топқа 5 үш орынды сандар (белгі-код) сәйкестендірілген: 113, 114, 115, 116, 117. Осы тәріздес, басқа топтама орындарында да өзіне тән сипаттамалар және соған сәйкес шартты белгі-кодтар *1-кестеде* көрініс тапқан.

Енді, осындай белгі-кодтар жиынын қалай пайдаланамыз,- деген сұрақ туындауы мүмкін. Бұл жайында зат есім сөздерге қатысты мақалада да қысқаша айтылған болатын. Етістік сөздерге қатысты *1-кестені* пайдалануда да айтарлықтай айырмашылық жоқ деуге болады. Осы тұста айта кетерлік жайт, топтама орындар санын (белгі-код орнын) тағайындау (белгілеу), біріншіден, зерттеу мақсатына сай, екіншіден, әр сөз табының лексика-морфологиялық, семантикалық ерекшеліктеріне байланысты деуге болады. Мысалы, зат есім сөздерге қатысты сегіз орын жеткілікті деп алсақ, етістік сөздерге қатысты *1-кестеде* ондай орындар саны төртеу-ақ. Ал басқа сөз табына қатысты немесе зерттеу мақсатына сай, топтау орын саны басқаша болуы да мүмкін.

Енді кестеде көрініс тапқан деректерді пайдалану жағын сөз етелік. Алғашында, зерттеуге алынған мәтін бөлігіндегі сөзқолданыстарды сөз табына ажырататын шартты белгілер қойылғаннан кейін, олардан компьютер арқылы қажетті жиілік сөздіктер алынуы керек.

Аталған жиілік сөздіктегі етістік сөзтұлғаларды (не басқа сөз табына қатысты сөздерді) бөліп алып, оларға *1-кесте* бойынша белгі-кодтарды сәйкестендіру қажет болады.

1-кесте **Етістік сөздерге лексика-морфологиялық ақпаратты сәйкестендіру бағдарламасы**

Белгі-код орны	**Етістік сөздердің лексика-морфологиялық сипаты**	**Белгі-код**
1	2	3
1	**Етістік сөз**	ЕТ/
2	**<u>Мағыналық және функционалдық жағынан топтастыру:</u>**	
	1) Амал-әрекет етістіктері Мысалы: *босат, көтер, күрес, қи, сыз, өлше* т.т.	100
	2) Қимыл-қозғалыс етістіктері Мысалы: *ауна, аудар, домала, секір, бүкірей* т.т.	101
	3) Қалып-сапа етістіктері. Мысалы: *жат, жантай, тұр, отыр, ұлғай* т.т.	102
	4) Ойлау-сөйлеу етістіктері. Мысалы: *айт, сөйле, де, ескер, жатта, ұмытпа* т.т.	103
	5) Өсу-өну етістіктері. Мысалы: *балала, жапырақта, гүлде, өс, қозыла* т.т.	104
	6) Бағыт-бағдар етістіктері. Мысалы: *бар, кел, қайт, әпер, әкет, қаш, түс, көтер* т.т.	105
	7) Көңіл-күй етістіктері. Мысалы: *жыла, қайғыр, өкін, күл, қуан, алақайла* т.т.	106
	8) Бейнелеу-еліктеу етістіктері. Мысалы: *жарқыра, күркіре, дүркіре, тарсылда* т.т.	107
	9) дыбыс-сес етістіктері	108
	10) көру-есту етістіктері.	109
	11) мінез-құлық етістіктері	110
	12) етістіктің тізімге енбеген түрлері	111
3	<u>Түбір етістік.</u> **Анықтама:** Етістік түбірі неше қилы формалар (есімше, көсемше, рай, етіс, шақ т.б.) тудыратын қосымшалардың барлығын алып тастағанда сақталатын бөлігі. **Мысалы: *аз, ал, айт, алда, ат, ас* т.б.**	112
	Есім негізді сөздерден жасалатын туынды етістіктер. Жұрнақтар арқылы жасалу жолдары: 1. ***–ла (-ле, -да, -де, -та, -те)*** жұрнағы арқылы жасалған туынды етістіктер: а) дене мүшелері атауларына жалғанады. Мысалы: *аяқта, жұдырықта, өкшеле*.	113
	Ә) Еңбек құралдары атауларына жалғанады. Мысалы: *арала, балтала, арқанда*.	114
	б) іс-әрекетті жүзеге асыруға объект болатын заттарға жалғанады. Мысалы: *жүнде, майла, тұзда, алтында.*	115
	в) Мекен я орын, өлшеу, дыбыс, қозғалыс, көрініс, төл және әлеуметтік,саяси, мәдени, тұрмыс, салт, табиғат т.т. атауларына жалғанады. Мысалы: *мекенде, өрле, төменде; аршында, метрле, тоннала; мыңқылда, шыңқылда; жарқылда, бұрқылда, ирелеңде; ботала, қозда; тәрбиеле, еркеле, жазала,қаумала, үймеле* т.т.	116

	г) Сын есім, сан есім, үстеу, еліктеуіш, одағай сөздерге жалғанады. Мысалы: *ақта, қарала, жаманда; онда, жүзде, екеуле; төменде, кейінде, ілгеріле; жымыңда, күлімде; аһла, уһле, ойбайла* т.т.	117
	2. ***–лан (-лен, -дан, -ден, -тан, -тен)*** жұрнағы арқылы жасалады. Мысалы: *ашулан, арлан, борышлан, иелен …*	118
	3. ***–лас (-лес, -дас, -дес, -тас, -тес)*** жұрнағы арқылы жасалған. Мысалы: *бірлес, көмектес, бәстес, достас, кезектес, сабақтас* т.б.	119
	4. ***–лат (-лет, -дат, -дет)*** жұрнағы арқылы жасалған. Мысалы: *дауылдат, тездет, түнделет* т.б.	120
	5. ***–а (-е)*** жұрнағы арқылы. Мысалы: *аша, деме, жаса, өрте, тіле, сына, міне* т.б.	121
	6. ***–ай (-ей, -й)*** жұрнағы арқылы. Мысалы: *күшей, мұңай, қартай, кеңей, көбей, молай* т.б.	122
	7. ***-қар (-ғар, -кер, -гер)*** өнімсіз жұрнағы арқылы. Мысалы: *басқар, ескер, аңғар, теңгер, ақыр, қақыр, жазғыр, ысқыр, кекір* т.б.	123
	8. ***-ар (-ер, -р)*** көне жұрнағы арқылы. Мысалы: *жаңар, тазар, ескір, өзгер* т.б.	124
	9. ***-ал (-ел, -ал, -іл, -л)*** қосымшасы арқылы. Мысалы: *жоғал, оңал, тарыл, тіріл, теңел* т.б.	125
	10. ***-ық (-ік)*** көне жұрнағы арқылы. Мысалы: *ашық, бірік, зарық, дәнік, өшік* т.б.	126
3	11. ***-сы (-сі)*** және ***-ымсы (-імсі)*** көне жұрнақтары арқылы. Мысалы: *батырсы, босаңсы, көлгірсі, үлкенсі, апамсы* т.б.	127
	12. ***-сын (-сін)*** жұрнағы арқылы. Мысалы: *адамсын, білгішсін, көпсін, жамансын* т.б.	128
	13. ***-сыра (-сіре)*** жұрнағы арқылы. Мысалы: *айсыра, әлсіре, қансыра, жетімсіре* т.б.	129
	14. ***-ыра (-іре)*** жұрнағы арқылы. Мысалы: *барқыра, бұрқыра, дүркіре, күркіре* т.б.	130
	15. ***-ырай, (-ірей)*** жұрнағы арқылы. Мысалы: *бақырай, шақырай, кішірей, шікірей,* т.б.	131
	16. ***-ы, -і; -шы, -ші; -ан, -ен, -ын, -ін, -н; -ырқа, -ірке, -ырқан, -іркен; -ына, -іне; -қа, -ке, -ға, -ге*** т.б. көне жұрнақтар арқылы. Мысалы: *байы, жасы, желпі, кеңі; аунақшы, дөңбекші; тасырқа, мүсірке, ашырқан, шіміркен; есіне, қатына, пысына; бүрке, иіске, қозға; басын, жирен, оян, үйрен* т.б.	132
	Етістік негізді сөздерден жасалатын туынды етістіктер	133
4	Күрделі етістіктер	140
	Жалаң етістік	141

Ескерту ретінде айта кететін жайт, зерттеуді жиілік сөздіксіз-ақ бірден мәтінмен де жүргізуге болады, бірақ бұл тәсілде қайталанатын мәтін бірліктері зерттеушінің көп уақытын алуы мүмкін. Қалай болғанда да зерттеу жұмысы тізбек құрайтын сөзқолданыс жүйесінен етістік сөзтұлғаларды (немесе басқа сөз табы) кезек-кезегімен бөліп алып қарастыруды қажет етеді.

Белгіленген бір орынға сол топтамадағы белгі-кодтар ішінен біреуі ғана таңдалып алынатынын үнемі есте сақтау қажет.

Мысал ретінде мына сөйлемде кездесетін етістік сөзтұлғаларды бөліп алып қарастырайық: *«Кешке әрі тоңып, әрі шаршап аһлап, үһлеп отырған, кемпір өзін өлтіріп кете жаздаған кім екенін есіне алды»* (С.Көбеев).

Етістік сөзтұлғаларға жататындар: *тоңып, шаршап, аһлап, үһлеп, отырған, өлтіріп кете жаздаған, алды.*

Енді *1-кестеде* көрсетілген деректер бойынша етістіктерге белгі-код мәндерін сәйкестендірейік:

тоңып	ЕТ/	106	133	141
шаршап	ЕТ/	106	133	141
аһлап	ЕТ/	106	133	141
үһлеп	ЕТ/	106	133	141
отырған	ЕТ/	102	133	141
өлтіріп	ЕТ/	100	133	141
кете жаздаған	ЕТ/	105	140	140
алды	ЕТ/	111	133	141

Сонымен, әрқилы етістік сөздер өздерінің жасалу және мағыналық ерекшеліктеріне қарай неше түрлі белгі-код мәндерін қабылдай алатынын аңғаруға болады.

Корпустық лингвистиканы жаңа білім аясы деп түсінуге байланысты «әдіс» ұғымының орны мен рөлі айтарлықтай өзгеріске ұшырауы мүмкін. Мәселен, корпустық лингвистиканы компьютерлік лингвистиканың бір бөлігі деп қарастыратын болсақ, онда төртінші топқа қатысты ақпараттану әдістері басымшылық етер еді. Ал В.П.Захаровтың жоғары оқу орындарында «Корпустық лингвистика» пәнін оқыту курсында мәтінге лингвистикалық шартты белгілер енгізуге, корпустарды құру технологиясына, корпустан қажетті ақпаратты іздеу әдістеріне (конкордансерілерге, корпустық менеджерлерге, сұрату тілдеріне) көбірек көңіл бөлінген [7; 8].

Мәтіндер корпусын қолдану аясы неше түрлі болғандықтан, олар лингвистикалық, филологиялық және педагогикалық салаларға топтастырылуы мүмкін. Ондай топтарға жататын пәндер: лексикография, грамматика, мәтін лингвистикасы, типология, контрастивтік лингвистика, аударма жасаудың теориясы мен практикасы, әдебиеттану ғылымы, лингводидактика. Қазіргі кезде корпустық лингвистиканың қолдану аясы бұрынғыдан да кеңейе түсуде.

Әдебиет

1. *Жұбанов А.Қ.* Корпустық лингвистика – қазақ тіл білімінің жаңа бағыты // Тілтаным. Языкознание. ҚР БжҒМ А.Байтұрсынұлы атындағы Тіл білімі институты, сәуір-мамыр-маусым. – Алматы.2009 жыл, №2 (34) – 3-11 бб.

2. *Жұбанов А.Қ.* Қазақ тілі мәтіндер корпусының компьютерлік базасын құрудың алғышарттары // ҚР ҰҒА-ның академигі, ф.ғ.д,, проф. Р.С.Сыздықтың 85 жылдығына орай өткізілген «Әдеби тіл және қазақ тілінің өміршеңдігі» атты ғылыми-теориялық конференцияның материалдары. –Алматы, 2009. –175-179 бб.

3. *Жұбанов А.Қ.* Қазақ әдеби тілінің электрондық корпусын түзудің теориялық бастаулары // Тіл және мәдениет: тілдің антропоөзектік парадигмасы.

Профессор Ж.А.Манкееваның 60 жылдығына арналған республикалық ғылыми-теориялық конференцияның материалдары. –Алматы: А.Байтұрсынұлы атындағы Тіл білімі институты, 2010. – 191-197-бб.

4. *Мырзабеков С.* Статистико-лингвистический анализ структуры глагола современного казахского языка. Автореф. дисс. … канд. филол. наук. Алма-Ата, 1973. С. 32.

5. *Жұбанов А.Қ.* Қазақ тілі корпусындағы зат есім сөздерге лексико-морфологиялық деңгейдегі белгіленім қою бағдарламасы // Қазақ тілінің тарихы мен түркітану мәселелері /Жауапты редактор ф.ғ.д., профессор М.Малбақов. –Алматы: А.Байтұрсынұлы атындағы Тіл білімі институты, 2011, – 228-233 бб.

6. *Ысқақов А.* Қазіргі қазақ тілі. Алматы: «Ана тілі», 1991. 384 б.

7. *Захаров В.П.* Корпусная лингвистика: Учебн.-метод. пособие. –СПб., 2005. –48 с.

8. *Захаров В.П.* Корпусная лингвистика [Электронный документ] // http://download.yandex.ru/class/zakharov/CL

ҚОРЫТЫНДЫ

Жинақта қамтылған мақалалар қазақ тілінің қолданбалы саласының мынадай өзекті мәселелерін айқындауға арналған деуге болады:

Қолданбалы тіл білімінің «статистикалық лингвистика» (квантитативті лингвистика) тарауы бойынша ғылыми ізденістер ғалым Қалдыбай Бектаевтың атымен және оның ғылыми мектебімен тығыз байланысты. Қазақ тілінің әртүрлі стильдерінің мәтіндері бойынша, алғашында электронды-есептеу машинасының (ЭЕМ), кейіннен компьютерлердің көмегімен алынған жиілік сөздіктер (әліпбилі жиілік, жиілік, кері әліпбилі) тілдің құрылымдық жағының статистикасы жайында көптеген мәліметтер алуға мүмкіндік туғызды. Осылайша, қазақ тіліндегі бай статистикалық мәліметтер негізінде бірнеше кандидаттық және докторлық диссертациялардың қорғалғаны көпшілікке мәлім. Бүгінгі таңда А.Байтұрсынұлы атындағы Тіл білімі институтының «компьютерлік лингвистика» бағыты бойынша қол жеткен ғылыми нәтижелері айтарлықтай деуге болады.

Статистикалық деректерді өңдеуде компьютерлік технологияны пайдалану, тілдік мәселелердің шешімін тез арада табу, математикалық статистика әдістерін басқа ғылымдар тәсілдерімен (ақпараттық теория, математикалық логика және т.б.) ұштастыру қазіргі кездегі отандық және шетелдік тіл білімінің даму бағытына сай келеді.

Тіл зерттеу тәжірибесінде сандық зерттеулердің қажеттігін математикалық лингвистика мамандарымен қатар дәстүрлі тіл білімінің аса көрнекті өкілдері де мойындап келеді. Олардың пікірінше, жазба әдебиет пен сөйлеу тіліндегі түрлі сөздердің әр ыңғайдағы қолданылу жиілігін анықтау тәсілі тілдің құрылымды-грамматикалық, кейде стильді-семантикалық айырымдарын ажыратумен қатар стильдерге тән қасиеттерді де анықтауға көмектеседі. Барлық грамматикалық категорияларды сандық тәсілмен талдау олардың әдеби тілдегі функционалдық салмағын нақты көрсетуге мүмкіндік береді. Қолданбалы лингвистиканы функционалдық көзқарас тұрғысынан қарастырсақ, оны тіл қызметінің ең қолайлы әдіс-тәсілдерін ұсынатын ғылыми пән деп те түсінуге болады. Тілдің коммуникативті функциясының қолайлы жақтарын ұсынатын пәндер – аударма теориясы, машиналық (компьютерлік) аударма, ана тілі мен шет тілдерін оқытудағы ақпараттық-іздестіру жүйесінің теориясы мен практикасы, ақпараттық жасанды тіл жүйесін жасаудағы таңбалау теориясы.

Ғылым мен техниканың дамуы тіл қызметінің қоғамдағы коммуникативтік қолданыс аясын кеңейте түскені бәрімізге мәлім. Тіл өзінің адамдар арасындағы қарым-қатынастық қызметін атқаруымен бірге адам мен компьютердің қазақ тіліндегі диалогын да жүзеге асыруы қажет деп білеміз. Осы себептен жақын болашақта табиғи тіл және оның түрлері ақпарат таратудағы **«адам–компьютер–адам»** аралығындағы коммуникативтік жүйе негізін құрайды деп сенім білдіруге болады.

Компьютерлік лингвистика информатиканың тілдік негіздемесін қарастыруымен қатар «тіл» мен «ойлау» арасындағы байланысты және одан

туындайтын тікелей болмысты компьютер арқылы модельдеуге қатысты мәселелермен шұғылданады. Компьютерлік лингвистиканы тілдік бірліктерге қатысты проблемалар да қызықтырады. Ондай бірліктер сөзден кіші, үлкен және оған тең де болуы мүмкін. Бұл жердегі сөзден үлкен деп отырғандарымыз – сөз тіркестері, сөйлемдер (айтылымдар) және толық мәтіндер. Аталып отырған салада тілді зерттеу кезінде қолданылатын амал-тәсілдерді модельдеу жолдары да қарастырылады. Мәселен, мәтін мазмұнын ашу жолдарын немесе аудармашының бір тілдегі мәтінді екінші тілге аудару жолдарын модельдеу мәселелері бұған мысал бола алады.

Қазақ тілінің компьютерлік саласы тиісті дәрежеде дамуын, оның қатаң ережеге бағынуынан, мәтін бойындағы тілдік бірліктерді тануға қажетті формалды қасиеттерінің айқындалуынан, омоним сөздер мен көпмағыналы сөздердің және синоним сөздер қатарын сирету мен реттеу мәселелерінің өз шешімін дұрыс табуынан байқауға болар еді. Қазақ тіл біліміндегі зерттеу кезеңдеріне, әсіресе, лингвистикалық ақпараттармен жұмыс істеу: оларды жинақтау, іздеу, сақтау, шет тілдегі ақпараттарды қазақ тіліне аудару, ақпараттық библиографиялық іздеу, қазақ тіліндегі мәтіннен реферат алу және т.б. істерге логикалық роботтың, яғни компьютердің араласуы жақын болашақта мемлекеттік тұрғыда қайта қаралып жатса, бұл мәселе өндірістік дәрежеге дейін көтерілуі де мүмкін.

Осы уақытқа дейінгі Қазақстандық «Статистика тобының» қазақ тілін компьютердің көмегімен лингво-статистикалық зерттеу барысындағы қол жеткізген көптеген ғылыми нәтижелерін, яғни әр мақсатпен түзілген жиілік сөздіктер мен олар арқылы алынған құнды деректер мен тұжырымдарды әрі қарайғы іс-тәжірибемізде дұрыс пайдалана білуіміз керек. Ең алдымен, сол ғылыми жетістігімізді үнді-еуропа тілдеріндегі «компьютерлік лингвистика» саласы бойынша орындалып жатқан ғылыми-зерттеу жұмыстарының деңгейіне дейін көтеру қажет деп білеміз.

Сонымен бірге, қазақ сөзінің «Тіл – қазына» атты деректер базасы үлкен көлемді мәтіндер мен көп көлемді сөздіктерге қатысты ақпараттық-сұраныс жүйесі түрінде көрініс табуы да қазақ мәтіндер корпусының компьютерлік базасын құрудың мақсат-міндетіне айналып отыр.

Көптеген ғалымдардың пікірлерін қорыта келе, біз қазақ сөзінің «Тіл – қазына» атты деректер базасы жүйесінің түсінігі – көп тармақты әрі терең автоматтандырылған жүйені аңғартады демекпіз және оның атқаратын міндеті ақпараттық барлау мен тіл білімі саласындағы (жалпы филология бағыттағы) зерттеушілік болуы тиіс.

Қазақ сөзінің «Тіл – қазына» атты деректер базасы зерттеуші адамның қалып-қабілетін әлденеше есе арттырумен бірге ізденушінің шығармашылық қуат көздерін аша түсетіні айғақ, сөйтіп бұл жаңа мүмкіндіктер қазақ тілінің жүйелілік қасиеттерін жетілдіруге және тіл жүйесін мұқият тануға жұмсалады деп тұжырымдауға әбден болады. Әрине, бұл аталып отырған қазақ сөзінің «Тіл – қазына» атты деректер базасын толық сипаттап, өмірге келтіру үшін оны мемлекеттік дәрежеде, яғни ҚР Білім және ғылым министрлігі, Ғылым комитеті, Тіл комитеті деңгейінде қаралуы қажет.

Себебі, мұндай жалпы филологиялық, ұлттық және мәдениет дәрежесіндегі күрделі жүйені іске қосу – қажетті мамандарды, мысалы, жоғары дәрежелі инженер-бағдарламашыларды, компьютерлік лингвистика саласының мамандарын және тиісті қаражат мүмкіндігін шешуді талап етеді.

Сөз соңында, қолданбалы тіл білімі қандай ғылым саласы деген сауалға жауап беру үшін оны нақты ғылымдар саласымен және табиғат заңдылығымен салыстыра келе, мынаны айтуға болар еді. Жаратылыстану және нақты ғылымдар уақыт көшімен дами отырып, табиғи ортадағы бәйтерекке ұқсап өседі: кейбір бұтақтары қурап түсіп жатса, кейбіреуі жайқалып желек жайып, дарақ діңінің жуандауына байланысты оның тамыры да тереңге – тарих қойнауына кетеді. Ал тіл білімінің және оның қолданбалы саласының дамуы бұдан өзгеше – бұл кең көгалдағы құлпырған гүлдердің өрнегі іспеттес, және де бұл көгал сыйқырға толы: жаңа гүлдер шыққанда ескілері солып қалмайды, өзінің айқын да балауса күйін жоғалтпайды. Сондықтан қолданбалы тіл білімінің жаңа саласының қазіргі кездегі дамуы бұрын қол жеткен нәтижелер негізінде «жұпар ииске» ие болуда. Гүл түрлерінің, хош иістерінің әртүрлі болатындығы сияқты мәтіндік бірліктердің мән-мағынасы мен мазмұнын лингвистикалық автоматта беру тәсілдері де әр бағытта дамып келеді. Солардың ішінде тілдік бірліктерді мәтін лингвистикасы тұрғысынан қарастыру жолы ғылымның соңғы жетістігі деуге боларлықтай. Осыған байланысты тілді зерттеуге динамикалық көзқараспен қарау, әсіресе, айтылымның туындауы мен әртүрлі түсінікке ие болу сипаты және оны тілдік қарым-қатынас тұрғысынан қарастыру нәтижелері аса құнды ғылыми жетістік болуда.

МАЗМҰНЫ

	Құрметті оқырман! (Алғы сөз ретінде)	3
	АЛҒЫ СӨЗ (Асқар Құдайбергенұлы Жұбанов – 75 жаста)	
1	**КІРІСПЕ** Компьютерлік лингвистика бөлімі	9
	І БӨЛІМ ..... **ҚАЗАҚ МӘТІНІНІҢ КВАНТИТАТИВТІ ҚҰРЫЛЫМЫ.** **(Квантитативная структура казахского текста).**	27
2	Частотные словари и информационные характеристики языка ……...	27
3	К вопросу об автоматизации лингвистических исследований	28
4	Автоматизация некоторых лингвистических процессов	30
5	Автоматическое построение частотных словарей (прямого и обратного) ……...	36
6	Универсальная программа получения на ЭВМ «Минск-22» алфавитно-частотного списка сочетаемости графем	42
7	Исследование статистической структуры тюркского текста …….	47
8	Становление статистического метода в казахском языкознании	48
9	Статистика лексико-морфологических форм прилагательных в романе М.Ауэзова «Абай жолы» («Путь Абая»)	53
10	Индексация и кодирование текстовой информации для ввода в ЭВМ	63
11	Задача получения на эвм частотных списков лингвистических единиц	69
12	Вероятностно-статистическое моделирование тюрских текстов	85
13	Вероятностные законы распределения классов слов казахского текста ……..	102
14	Статистическое исследование казахского текста с применением ЭВМ (на материале романа М.Ауэзова «Абай жолы»). Автореферат диссертации на соискание ученой степени кандидата филологических наук …….	107
15	Кибернетика и лингвистика	133
16	Информационно-статистическая типология тюркского текста ……	136
17	К вопросу о графемной статистике казахского текста ……..	138
18	О ходе подготовки к составлению словаря языка М.О.Ауэзова (функции ЭВМ)	142
19	Преобразование с помощью ЭВМ казахских текстов в фонетическую запись …	143
20	Состояние и перспективы исследования тюркских языков с применением статистико-инжерных методов …….	145
21	Использование ЭВМ в авторской лексикографии ……..	146
22	Статистический анализ грамматических оформителей на ЭВМ ……..	148
23	О создании машинного фонда казахского языка ……	149
24	О проблеме создания машинного фонда тюркских языков ……...	151
25	Создание словарной базы данных обще– и межтюркских моносиллабов: к проблеме разработки машинного фонда тюркских языков	159
26	Алгоритмизация задачи получения на ЭВМ фонетической транскрипции казахского текста ……	161
27	О лингвистических исследованиях с применением вероятностно-статистических методов в тюркских языках …….	162
28	Использование ЭВМ для выделения опорных слов текста	173
29	Автоматический отбор опорных слов текста в целях оптимизации обучения казахскому языку ……..	174
30	Автоматический анализ морфологии казахской словоформы …….	175
31	Түркі тілдерінің машиналық қорын жасаудың алғышарттары	179

32	Текст лингвистикасы жайында	188
33	Аормализация процедуры построения текстов естественного языка	195
34	М.Әуезов тілінің сөздіктерін жасау тәжірибесі	200
35	Профессор Қ.Жұбанов және қазақ тілін зерттеудің кейбір қолданбалы тәсілдері	208
36	Қолданбалы тіл білімі	212
37	Машиналық аударма	213
38	Проф. Құдайберген Жұбановтың буын теориясы және оның компьютерлік лингвистикадағы орны	214
	II БӨЛІМ ……..... **ҚАЗАҚ МӘТІНІНІҢ МАЗМҰНЫНЫН ФОРМАЛДАУ. (Формализации содержания казахского текста).**	219
39	Статистическое исследование казахского текста на уровне абзацев …….	219
40	К проблеме семантической классификации знаменательных классов слов казахского языка …….	226
41	Основные принципы формализации содержания казахского текста …….	238
42	Формальное описание словарной статьи автоматического словаря казахского языка (АСКЯ) ……...	240
43	Основные принципы формализации содержания казахского текста. Автореферат диссертации на соискание ученой степени доктора филологических наук	249
44	Компьютерлік лингвистика және қазақ тіл білімі	301
45	«Адам-компьютер-адам» жүйесіндегі табиғи тіл қызметінің ерекшеліктері	310
46	Абзац – мәтін құрылымының негізгі семантикалық компоненті	319
47	Тілді логикалық талдау бағытының кейбір мәселелері	326
48	Қазақ мәтінін абзац деңгейінде зерттеудегі мазмұнға қатысты тірек сөздерді статистикалық жолмен анықтау	335
49	Қазақ мәтінінің ықтималды-статистикалық моделін құру маңызды мәселе	341
50	Қолданбалы лингвистика пәні және оның негізгі бағыт-бағдары	348
51	Мәтін абзацтарын заттық-логикалық мазмұн тұрғысынан зерттеу	355
52	Құдайберген Жұбанов және қазақ тіл біліміндегі статистикалық әдістің жетістіктері	362
53	Қазақ тілтанымындағы статистикалық әдістің орны	370
54	Қазақ мәтінін талдау мен жинақтаудың автоматты жүйелерін құру мәселесі	382
55	Оқу лексикографиясының ғылыми пән ретіндегі ерекшелігі	387
56	Ахмет Жұбанов зерттеулеріндегі қазақтың күй жанры мен домбыра музыкалық аспабы туралы	396
57	Лексикография саласындағы автоматтандыру тәсілі	400
58	Құдайберген Жұбанов зерттеулеріндегі модельдеу әдісінің көрінісі	408
59	Ғылыми-техникалық лексикография және арнаулы сөздіктер түрлері ……	414
60	Автоматтанған оқу жүйесінің негізгі ұстанымдары	423
61	Табиғи тілдің мәтін морфологиясын автоматты түрде талдау мәселесі	429
62	Машиналық аударманың ғылыми бағыт ретіндегі даму кезеңдері	434
63	Баспа ісін автоматтандыру әдісіндегі лингвистикалық аспектілер	441
64	Қазақ жазба мәтіндеріндегі графема қолданысының статистикасы	447
	III БӨЛІМ ..... **ҚАЗАҚ ТІЛІНІҢ «ТІЛ – ҚАЗЫНА» АТТЫ ҰЛТТЫҚ КОМПЬЮТЕРЛІК ҚОРЫ. (Национальная компьютерная база казахского языка «Тіл – қазына»).**	453
65	Қазақ тілінің «Тіл – қазына» атты ұлттық компьютерлік қорының мүмкіндіктері жайында	453

66	Жаңа ақпараттық технологияға көшу жағдайындағы қазақ лексикографиясы	460
67	Автоматты (машиналық) аударма	468
68	Қазақ тілінің «Тіл – қазына» атты ұлттық компьютерлік қорын құрудың алғышарттары	494
69	Есептеуіш лексикография – қазақ тіл білімінің жаңа ғылыми бағыты	499
70	Проф. Қ. Жұбанов және қолданбалы тіл біліміндегі формалды модельдер	505
71	Проф. Қ. Жұбанов зертттеулері – қолданбалы тіл білімінің жаңа саласына ғылыми бағыт-бағдар	510
72	Проф. Қ. Жұбанов және қолданбалы тіл білімінің жаңа бағыты	516
	IV БӨЛІМ ..... **ҚАЗАҚ ТІЛІНІҢ ҰЛТТЫҚ МӘТІНДЕР КОРПУСЫН ҚҰРУ. (Создание национального корпуса текстов казахского языка).**	523
73	Қазақ тілі мәтіндер корпусының компьютерлік базасын құрудың алғышарттары	523
74	Корпустық лингвистика – қазақ тіл білімінің жаңа бағыты	530
75	Автоматтанған ақпараттық-іздестіру жүйесінің кейбір ұғымдары мен тілі	538
76	Қазақ тілінің «Тіл – қазына» атты компьютерлік базасы жайында	545
77	Қазақ әдеби тілінің электрондық корпусын түзудің теориялық бастаулары	552
78	Мәтінді автоматты өңдеу жүйесінің құрылымы жайында	561
79	Мәтінді автоматты өңдеу жүйесіндегі лингвистикалық ақпарат базасы	567
80	Қазақ сөзінің «Тіл – қазынасы»	573
81	The founder of the kazakh linguistics and the victim stalin's repressions (tragic fate of Professor Khudaibergen Kuanovich Zhubanov)	575
82	Мемлекеттік тілдегі ұлттық «Тіл – қазына» деректер базасының теориялық негіздері	584
83	Ұлттық филологиямыздың тарихында алғаш профессор атағы берілген құдайберген қуанұлы жұбановтың тағдыры жайында	613
84	Түркі тілдерінің көпаспектілі компьютерлік қорын құру қажет	622
85	Қазақ лексикографиясының «Тіл – қазына» атты электрондық базасы жайында	632
86	Қазақ тілі корпусындағы зат есім сөздерге лексика-морфологиялық деңгейдегі белгіленім енгізу бағдарламасы	636
87	Қазақ тілінің автоматтанған компьютерлік базасын құру – заман талабы	641
88	Қазақ тілінің аннотацияланған мәтіндер корпусындағы сын есім сөздердің лексика-морфологиялық құрылымына белгі-код қою бағдарламасы	652
89	Семантические классы существительных, использованных в модели порождения казахского текста	658
90	Тұрақты сөз тіркестерін кодтау және модельдеу мәселесі	665
91	Қазақ тілінің аннотацияланған мәтіндер корпусындағы етістік сөздерге лексика-морфологиялық белгі-код қоюдың (белгіленім) алғышарттары	668
92	ҚОРЫТЫНДЫ	677