ӘОЖ 81.33
ХҒТАР 16.01.29
DOI10.48371/PHILS.2022.66.3.010

ҚАЗАҚ ТІЛІНІҢ ҰЛТТЫҚ КОРПУСЫНДАҒЫ ЕТІСТІКТЕРДІҢ ЛЕКСИКА-СЕМАНТИКАЛЫҚ БЕЛГІЛЕНІМ ӘЗІРЛЕМЕСІ: ӘЛЕМДІК ТӘЖІРИБЕ, КЛАССИФИКАЦИЯЛАУ, КОРПУСТА БЕЛГІЛЕУ

*Момынова Б.Қ.[1], Имангазина М.А.,[2] Анесова У.Г.[3]
*[1]фил.ғ.д., профессор, А. Байтұрсынұлы атындағы Тіл білімі институты, Алматы, Қазақстан,
[2]1-курс докторанты, А. Байтұрсынұлы атындағы Тіл білімі институты, Алматы, Қазақстан,
[3]PhD доктор, әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,
*[1]e-mail: momynova_b@mail.ru,
[2]e-mail: meruyertimangazina@mail.ru,
[3]e-mail: a_ymit@mail.ru

Аңдатпа. Мақалада әлемдік корпус жасау тәжірибесіндегі негізгі белгіленімнің бірі лексика-семантикалық белгіленім әзірлемесін жасау мәселесі қарастырылады. Нақтырақ айтқанда, компьютерлік лингвистикаға және лексика-семантикалық классификацияға қатысты отандық және шетелдік ғалымдар еңбектеріне шолу жасалады, қазақ тілінің Ұлттық Корпусындағы етістіктердің лексика-семантикалық белгіленімін жасау кезеңдері көрсетіледі, практикалық негізі түсіндіріледі.

Ақпараттық технологиялардың жедел қарқынмен дамуы тіл білімі зерттеушілерінен электронды ресурстарды меңгеруді талап етуде. Тілді бағдарламалауды зерттейтін және жүзеге асыратын тіл білімінің саласы – корпустық лингвистика. Қазақ тілінің Ұлттық корпусын жасау тілдің әр деңгейлері бойынша автоматты түрде талдау жасайтын белгіленімдер жасауға негізделеді. Сөздерді лингвистикалық аннотациялаудағы күрделі белгіленімнің бірі – лексика-семантикалық белгіленім. Орыс, қалмақ және т.б. тілдер корпусымен салыстырғанда Қазақ тілінің Ұлттық корпусындағы лексика-семантикалық белгіленім сөз мағынасына, яғни семаға тереңірек бойлайды. Сол себепті етістіктің кіші (жеке) лексика-семантикалық топтарының саны – 72. Бұл қолданушыға өзіне қажетті ақпаратты нақтырақ табуға мүмкіндік береді. Белгіленім жүйесінің қолдану интерфейсі кез келген қолданушыға, осы істің маманына да, енді үйреніп жатқан басқа саладағы мамандарға да жеңіл, түсінікті болуы көзделеді, лексика-семантикалық топтар қысқа әрі нақты атаулармен беріледі.

Корпус базасына енгізуге 18 200 етістік жинақталып, мағыналық реңктері зерттелуде. Зерттеу барысында етістіктерге лексика-семантикалық белгіленім әзірлеуді бес түрлі сипаттама бойынша беру ұсынылды. Бірінші: сөзжасамдық сипатына қарай дара, күрделі; негізгі, туынды; екінші: лесика-грамматикалық категориялар негізінде салт, сабақты; болымды, болымсыз; коннотативтік сипатына қарай жағымды, жағымсыз, бейтарап деп жіктеледі. Етістіктердің мағынасын

тереңірек ашу үшін ортақ және айырушы семаларына байланысты іштей ірі (лексика-семантикалық) және кіші (семантикалық) топтарға бөлінеді.

Мақала BR11765619 «Мемлекеттік тілдің ақпараттық-инновациялық базасы ретіндегі Қазақ тілінің Ұлттық корпусын әзірлеу: ғылыми-зерттеу және оқыту интернет ресурсы» атты зерттеу жобасының аясында жазылды.

Тірек сөздер: корпустық лингвистика, белгіленім, лексика-семантикалық классификация, етістік, семантика, мағына, категория, лингвистикалық аннотация.

Негізгі ережелер

XX ғасырдың басында қазақ тілінің оқулықтары жазылып, жеке сала қалыптаса бастады. Ал XX ғасырдың ортасына қарай әр саласы академиялық ғылыми зерттеу нысанына айналды. Одан бергі кезеңде тіл грамматикасы, лексикасы, фонетикасы, тарихы динамикалы зерттелді. Ал Тәуелсіздік алғаннан кейін қазақ тіліндегі жаңа кезең ерекшелігі неде деген сұрақ туындайды. Бұл өткенді саралау, сараптау және дамытудан басталып, әлемдік лингвистикадағы жаңа білімдер жүйесін меңгеру, әр тақырыпқа тереңірек бойлауды жүргізу қолға алынды. Яғни тіл дамуын қоғам дамуымен сабақтастыра отырып, жаңа заман жетістіктерін игеріп, меңгеріп отыру талап етіледі. Корпустық лингвистиканың қазақ тіл біліміндегі орнын осындай жаңару, жаңа технологияларды ұтымды пайлаланудағы басты жетістік ретінде атап айту қажет.

Инновациялық зерттеулер қазақ тіл білімінде Кеңес Одағы кезінде-ақ Қазақ ССР Ғылым Академиясының Тіл білімі институтында 1969 жылы өткен «Түркі тілдеріндегі статистикалық және ақпараттық зерттеу» атты Бүкілодақтық ғылыми жиыннан бастау алды. 1970 жылы математикалық статистика саласының маманы, белгілі математик-ғалым Қ. Бектаевтың жетекшілігімен Институтта «Тіл статистикасы және автоматтандыру» ғылыми тобы құрылады. Бұл ғылыми топ алғашқыда қазақ тілінің әртүрлі стильдері бойынша жиілік сөздіктер жасаумен айналысты. Мәселен, өткен ғасырдың 90-жылдары М.О.Әуезовтің 20 томдық шығармалар жинағы ЭЕМ жадына енгізіліп, нәтижесінде 1995 жылы «М.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері» жарық көрді. Одан беріде компьютерлік лингвистика, тілдік бірліктерді формальды модельдеу бағытында жұмыстар жүргізілді. Әлемдік тіл біліміндегі лингвистикалық ғылыми зерттеулерді цифрландыру, оңтайландыру, автоматтандыру үдерісі қазақ тіл біліміндегі қолданбалы лингвистика, әлемдік лингвистикадағы жаңа бағыт – корпустық лингвистика саласын зерттеу мәселесін көтерді.

Кіріспе

Корпустық лингвистика әр ұлт тіліндегі мәтіндерді компьютерлік базаға жинақтап, оны программалық басқару жүйесіне салатын Ұлттық корпустарды құрастыру мен оны пайдалану мәселелерін зерттейді. Әлемдік тіл білімінде корпус жасау ісі 1960 жылдан Браун корпусынан басталады. Өзге тілдердің көбінде қазірде Ұлттық корпустар жасалған. Корпус жасау қазіргі ақпараттық технология заманында заманауи

инновациялық құрал жасау деген сөз. Ұялы телефондарға орнатылған сансыз электрондық қызметтер орын алған бүгінгі жағдайда ғылыми зерттеу құралдарын автоматтандыру, жас зертеушілерге тиімді әрі қолжетімді ету, ғасырлар бойы жинақталған түрлі стильде жазылған жазба мәтіндерді, құнды шығармаларды т.б. рухани мұрамызды компьютерлік жадта сақтау – өзекті мәселе. Қазақ тілінің Ұлттық корпусын жасауға қатысты жұмыстар қазіргі таңда Тіл білімі институтында қарқынды жүруде.

Қазақ тілінің Ұлттық корпусының құрылымы, жүйесі, ондағы мәтіндер мен метабелгіленімдерге қатысты жазылған монографиялар, ғылыми мақалалары аз емес. Алайда әлі де тереңірек қарастыруды қажет ететін тұстары да бар. Соның бірі – корпустағы лексика-семантикалық белгіленімді жүзеге асыру мәселесі.

Материалдар мен әдістерді сипаттау

Корпустық лингвистиканы тереңірек түсіну үшін әлем тілдеріндегі корпустар назарға алынды, әр корпустың негізгі бағыттары, жұмыс істеу жүйесі анықталды. Аталған жұмыстар салыстыру, талдау негізінде жүзеге асырылды. Қазақ тілінің Ұлттық корпусының құрылысы сипаттама, яғни синхрондық әдіс негізінде түсіндірілді. Мақалада қозғалып отырған мәселе – етістіктердің лексика-семантикалық белгіленіміне қатысты теориялық ақпараттар мен практикалық талдаулар құрылымдық әдіс арқылы зерттеліп, ұсынылды. Әр тұжырым лингвистикалық әдістерге сүйене отырып жасалып, мысалдармен дәлелденді. Қазақ тілінің бай лексикалық қоры, әсіресе қазақ тіліндегі етістіктерді зерттеуде назарға алынды.

Әдебиеттерге шолу

Корпустық лингвистика туралы жалпы зерттеулер әлемдік тіл білімінде Е.Финеганның «Language: its structure and use», Мак Энери және Э.Вильсонның «Corpus Linguistics» еңбегінде, В.Хахаровтың «Корпусная лингвистика» оқу құралында, А.Н.Барановтың «Корпусная лингвистика», «Ввдение в прикладную лингвистику», А.В.Зубов пен И.И.Зубованың «Информационные технологии в лингвистике» еңбектерінде көрініс табады. Сондай-ақ әр тілдегі Ұлттық корпустар туралы жеке талдаулар З.А.Сиразитдиновтың «Моделирование грамматики башкирского языка», В.З.Демьяниковтың «Англо-русские термины по прикладной лингвистике и автоматической переработке текста», Ч.Ф.Мейердің «English Corpus Linguistics», М.Кито мен М.А.Риссаненнің «A Language in transition: the Helsinki Corpus of English texts» зерттеулерінде жасалады.

Ал қазақ тіл біліміндегі компьютерлік лингвистиканың зерттелуі А.Жұбановтың 2002 ж. қорғаған, «Основные принципы формализации содержания казахского текста» атты қазақ қолданбалы лингвистикасына қосылған елеулі үлес саналатын докторлық монографиясынан басталады. Сондай-ақ А.Жұбановтың «Қолданбалы тіл білімінің мәселелері», «Қазақ

тексінің статистикасы» еңбектері, А. Жаңабекованың «Қазақ тілі мәтіндеріне морфологиялық белгіленім қоюдың ғылыми-тәжірибелік негіздері», «Аннотацияланған корпустардың ғылыми-практикалық маңызы», «Тілдік корпустардың ғылыми-зерттеу әлеуеті», Б.Карбозованың «Қазақ тілі мәтіндер корпусындағы омонимдер мәселесі», «Корпустық лингвистикадағы метабелгіленім ұғымы туралы түсінік», С.Құлмановтың «Қазақ тілі мәтіндері корпусының электрондық базасы жайында», «Корпусқа енгізілетін мәтіндердегі сөздерді морфологиялық талдау және оларға шартты белгілер қою принциптері», «Мәтіндер корпусы негізінде қазақ тілінің грамматикалық сөздігін түзудің практикалық аспектілері» сынды ғылыми мақалаларында аталған тақырып әр қырынан талданды.

Нәтижелер

Орыс тілінде «разметка» деп аталатын корпус термині қазақ тіл білімінде «белгіленім» деп аталады. («Белгіленім» термині – А.Жұбановтікі). Корпус – компьютерлік жадқа енгізілген мәтіндер жиынтығы ғана емес, сонымен бірге автоматты түрде түрлі лингвистикалық талдаулар жасайтын интерфейспен жабдықталған құрылғы. Корпуста тілді компьютерге бағындыру мақсатында тілдің әр деңгейі бойынша автоматты түрде талдау жасау міндеті алға қойылады. Автоматты талдауға, әсіресе, агглютинативті тілдер ыңғайлы. Өйткені мұндай тілдерде түбір мен қосымшаның жігі анық көрініп тұрады. Ұлттық корпустардың барлығында осылай түбір мен қосымшаны автоматты талдайтын морфологиялық белгіленім программасы алғаш жасалған. Қазақ тілінің Ұлттық корпусында да морфологиялық талдауыш (анализатор) қазақ тіліндегі сөз формаларды автоматты түрде түбір мен қосымшаға бөліп, олардың сөз табына қатысын және қосымшалардың грамматикалық сипаттамасын шартты белгілер арқылы көрсете алады. Мысалы, «балаларға» сөзін морфологиялық анализатор төмендегіше талдайды:

Балаларға

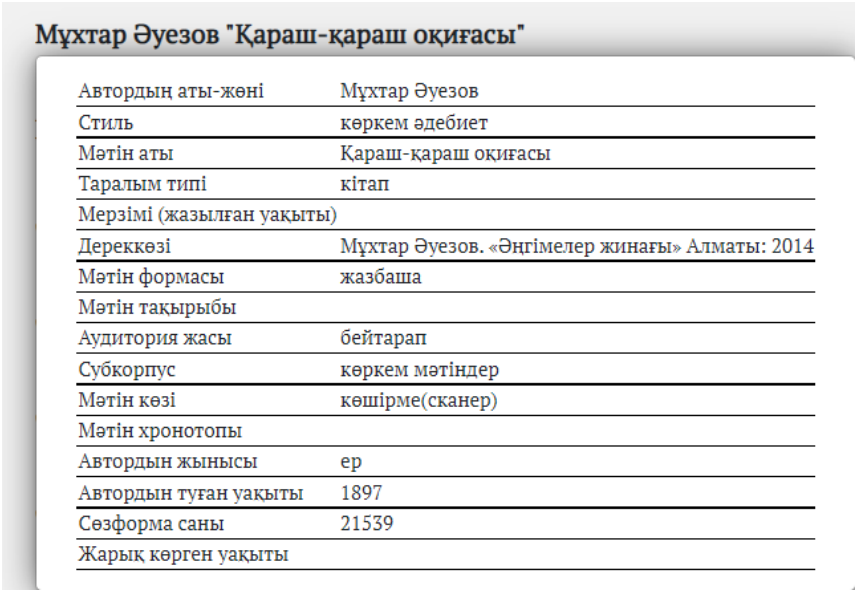
Лемма	бала
Морфология	зт,лар/кж+ға/бс

1-сурет. «Балаларға» сөзінің морфологиялық белгіленімі

Корпуста морфологиялық талдау ғана емес, тілдің басқа деңгейлеріне жасалатын тілдік талдаулардың барлығы қамтылуы тиіс. Мұндай жан-жақты лингвистикалық талдауды *«терең аннотацияланған»* корпус жасайды. Белгіленім түрлерін тіл деңгейлеріне сәйкес бөлуге болады: *морфологиялық белгіленім, синтаксистік белгіленім, метамәтіндік белгіленім, сөзжасамдық белгіленім, лексика-семантикалық белгіленім, анафоралық белгіленім.*

Метамәтіндік белгіленім – мәтін туралы берілетін ақпарат (мәтін авторы, атауы, жылы, стилі, жанры т.б.) экстралингвистикалық

белгіленім деп аталады. Ал тіл деңгейлерінің лингвистикалық белгіленімдері интралингвистикалық белгіленімге жатады. Қазақ тілінің Ұлттық корпусында метамәтіндік белгіленім 23 параметрмен жұмыс істейді.



Мұхтар Әуезов "Қараш-қараш оқиғасы"

Автордың аты-жөні	Мұхтар Әуезов
Стиль	көркем әдебиет
Мәтін аты	Қараш-қараш оқиғасы
Таралым типі	кітап
Мерзімі (жазылған уақыты)	
Дереккөзі	Мұхтар Әуезов. «Әңгімелер жинағы» Алматы: 2014
Мәтін формасы	жазбаша
Мәтін тақырыбы	
Аудитория жасы	бейтарап
Субкорпус	көркем мәтіндер
Мәтін көзі	көшірме(сканер)
Мәтін хронотопы	
Автордын жынысы	ер
Автордын туған уақыты	1897
Сөзформа саны	21539
Жарық көрген уақыты	

2-сурет. Метамәтіндік белгіленім үлгісі

Лексика-семантикалық белгіленім – сөздердің тақырыптық немесе лексика-семантикалық тобын көрсететін автоматты бағдарлама. Әлемдік корпустарда лексика-семантикалық белгіленім морфологиялық белгіленімге қарағанда кейінірек қолға алынған. Бұл аталған белгіленімнің күрделілігімен түсіндіріледі. Дегенмен Орыс тілінің Ұлттық корпусында семантикалық белгіленімді жасау үлкен қиындықтар тудырмаған. Себебі орыс тілінің сөздік құрамына негізделген 10 000 сөзден тұратын «Лексикограф» мәліметтер базасы 1992 жылдан жұмыс істеп тұр. Аталған жүйе Е.В. Падучеваның жетекшілігімен жасалған және ондағы етістіктер базасы аса бай. «Лексикографтың» артықшылығы – сөздерді семантикалық ерекшелігіне қарай талдауға мүмкіндік беруі. Жүйенің негізгі бағытын жоба авторлары былай көрсетеді: *«Лексикограф» жүйесі лингвистикалық зерттеудің жоғары технологиялық құралы ретінде жасалған. Оның негізгі мақсаты – алдын ала берілген семантикалық параметрлер (ең алдымен таксономиялық) бойынша сөздердің тізімін алу және алынған лексикалық кластардың тілдік сипатын, нақтырақ айтқанда, тіркесімділігін, басқару модельдерінің түрлерін, лексикадағы жүйелі қатынасқа негізделген диатетикалық қозғалысын зерттеу»* [11, 156 б.]. XX ғасырдың соңында жұмысын бастаған «Лексикограф» базасы корпус құрастырушыларға лексика-семантикалық белгіленім жасауда көп көмегін тигізді. Нәтижесінде Орыс тілінің Ұлттық корпусында *лексика-семантикалық іздеу* 2004 жылдың қазан айынан бастап іске қосылған: *«Көлемінің шағындығына қарамастан, теориялық тұрғыдан «Лексикограф» мәліметтерді эксперименталды зерттеу базасының лексика-семантикалық корпусты құрудағы рөлі зор және корпустың лексика-*

семантикалық сөздігінің толық негізін құрады деуге де болады» [11, 157 б.]. Орыс тілінің Ұлттық корпусында лексика-семантикалық белгіленім жасауда сөздерге негізгі үш белгісі бойынша сипаттама берілген. Олар: *1. Деривациялық сипаты бойынша:* «диминутив», «аугментатив», «аттенуатив», «nomen agentis», «nomen femininum» т.б.; *2. Лексика-грамматикалық категориялар бойынша:* сапалық, қатыстық (сын есім), деректі, дерексіз (зат есім) т.б.; *3. Жеке семантикалық сипаты бойынша:* тақырыптық (таксономиялық) класс, «бағалау», каузативтілік (етістік және етістік негізді сөздер) т.б.

Семантикалық белгіленім түрлері, ондағы лингвистикалық ақпарат әр сөз табының ерекшелігіне қарай беріледі. Мысалы, етістіктерді төрт түрлі белгісі бойынша аннотациялайды:

1. *Лексика-семантикалық топтар*, яғни мағынасына қарай *қозғалыс, физикалық әсер ету, жасау, жою, эмоция, сөйлеу* сынды семантикалық топтарға жіктейді. Белгіленімде әр лексика-семантикалық топ ағылшынша атаулармен белгіленеді. Мысалы:

t:move — движение (*бежать, дергаться, бросить, нести*)
t:move:body — изменение положения тела, части тела (*согнуть, нагнуться*)
t:put — помещение объекта (*положить, вложить, спрятать*)
t:impact — физическое воздействие (*бить, колоть, вытирать*)

2. *Каузативтілік.* Төмендегідей *каузативті* немесе *каузативті емес* дегенді білдіретін белгі қойылады:

ca:caus — каузативные глаголы (*показать, вертеть*)
ca:noncaus — некаузативные глаголы (*видеть, вертеться*)

3. *Көмекші етістіктерді сипаттау*. Олар *фазалық* және *каузативті көмекші етістіктерге* жіктелген:

aux:phase — фазовые (*начать, продолжать, прекратить*)
aux:caus — служебные каузативные (*вызвать, привести (к)*)

4.*Сөзжасамдық белгілеуде* етістіктер *префиксті етістіктерге, семельфактив* және *екіншілік имперфективке* ажыратылады:

d:pref — приставочные глаголы (*забегать, оглядеть*)
d:semelf — семельфактивы (*кивнуть, чихнуть, боднуть, качнуться*)
d:impf — вторичные имперфективы (-ива-, -ва-, -а-) (*выпивать, вбивать*).

Орыс тілінің Ұлттық корпусында 6 сөз табының лексика-семантикалық белгіленімі бар: *зат есім, сын есім, сан есім, етістік, есімдік, үстеу.* Зат есім жеке сөз табы түрінде емес, деректі, дерексіз, жалқы деп іштей үш түрге бөлініп беріледі. Тақырыптық топты «таксономия» деп атайды. Семантикалық фильтрлерді таңдауда бірнеше лексика-семантикалық топтарды қатар белгілеу бір сөздің бойындағы түрлі мағыналық реңкті ескеруге мүмкіндік береді. Ал етістіктер таксономия, каузативтілік, көмекші етістіктер және сөзжасам деген төрт негізгі белгісі бойынша сұрыпталып, 18 ірі және 9 кіші лексика-семантикалық топқа жіктелген.

№	*Жалпы семантикалық топтар*	*Жеке семантикалық топтар*
1	Қозғалыс	Дене қалпының, дене бөлігінің өзгеруі
2	Объектінің орналасуы	
3	Физикалық әсер	Физикалық объект құру
		Жою
4	Күйдің немесе белгінің өзгеруі	
5	Болмыс	Тіршілік
		Тіршілік басы
		Тіршілікті тоқтату
6	Объектінің орналасқан жері	Дененің кеңістіктегі орны
7	Байланыс және қолдау	
8	Меншік саласы	
9	Танымдық сала	
10	Қабылдау	
11	Психикалық сала, оның ішінде	Эмоция
		Ерік
12	Сөз	
13	Мінез-құлық адам	
14	Физиологиялық сала	
15	Табиғи құбылыс	
16	Дыбыс	
17	Түсі	
18	Иісі	

1-кесте. Орыс тілінің Ұлттық корпусындағы етістіктің лексика-семантикалық топтары

Қазақ тілінің Ұлттық корпусында етістіктерге лексика-семантикалық белгіленім жасауда орыс тілінің тәжірибесі ескерілді, ондағы әр сипаттаманың қазақ тілі етістіктерінің семантикасын ашуға қабілеттілігі назарға алынғанмен, каузативтілік және көмекші етістіктерге қатысты белгілеу қазақ тіліндегі етістіктерге сәйкес келмейтіні анықталды. Қазақ етістіктерінде каузативтілік семантикалық категориялардың да, лексика-грамматикалық категориялардың да қатарына жатпайды, яғни семантикаға емес, грамматикалық құрылымға негізделеді. «*Мәжбүрле», «күште», «көндір», «душар ет»* секілді таза каузативті етістіктердің саны көп емес, керісінше, көбінесе каузативтілік *кел-тір, ту-ғыз, жет-кіз* т.б. түрінде өзгелік етіс формалары арқылы көрінеді. Сондықтан каузативтілік «белгі-код» ретінде алынбайды. Префикс арқылы жасалатын етістіктер де қазақ тіліне тән болмағандықтан, көмекші етістіктерге қатысты сипаттама белгіленімге енбейді. Аталған белгілеулердің орнына етістіктер Ұлттық корпуста *салт/сабақтылық* және *болымды/болымсыздық* категориялары бойынша сипатталады. Бұл Қазақ тілінің Ұлттық корпусындағы лексика-семантикалық белгіленімнің негізгі ұстанымымен байланысты. *Қазақ тіліндегі лексика-семантикалық белгіленімнің мақсаты* – семантикаға басымдық беру және семантиканы

басты критерий етіп алу арқылы етістіктердің мағыналық реңкін жан-жақты ашып көрсету, мәнін түсіндіру. Себебі семантика сөз мәнінің, синтактикасының, қолданысының өзегі болып табылады.

Орыс тілінің Ұлттық корпусының үлгісімен Ресей Ғылым Академиясында Қалмақ гуманитарлық зерттеу институты Қалмақ тілінің Ұлттық корпусын әзірлеген. Қазіргі таңда қалмақ тілінің екі корпусы бар: біріншісі – РМГУ аспиранты Э.Ванькаева құрастырған «Қалмақ корпусы», екіншісі – «Қалмақ тілінің Ұлттық корпусы». Қалмақ тілінің Ұлттық корпусындағы лексика-семантикалық белгіленім төрт түрлі бағыт бойынша метасипаттама береді: *1. Лексика-грамматикалық сипаттама, 2. Лексика-тақырыптық сипаттама, 3. Бағалауыштық сипаттама, 4. Деривациялық сипаттама.*

Орыс корпусынан Қалмақ корпусының айырмашылығы – тақырыптық топтарды барлық сөз таптарына ортақ беруі. Корпус құрастырушылары бұлай берудің мәнін былай түсіндіреді: *«Біз де барлық сөз таптары үшін тақырыптық топтар ортақ деген пікірге қосыламыз. Біздің ойымызша, сөз семантикасы тілдегі константа болып табылады. Мысалы, жануарлар тақырыбына қатысты сөздер тобы кез келген тілде бар және әртүрлі сөз таптарындағы лексикалық бірліктерге ортақ сема бола алады. Оған қоса қалмақ тілінде изафеттік құрылымдар (модн гер «ағаш үй») кең тараған. Егер екі зат есім қатар келсе, онда оның алғашқысы сөз табы жағынан сын есім болады. Модн «ағаш» сын есім де, зат есім де бола алады. Қалай келсе де «t:plant» (тақырыбы: өсімдік) деп белгіленеді»* [12]. Барлық сөз таптарын ортақ тақырыптық топтарға жіктеу бәрін стандарт дайын үлгіге салу арқылы техникалық жұмысты жеңілдетеді, жеке зерттеуді қажет етпейді, бірақ «стандартты топтастыруда» сөз мағыналарын, мағыналық реңктерін ашу қиын. Сол үшін әр сөз табының лексика-семантикалық белгіленімі жеке қарастырылып, классификациялары бөлек жасалғаны жөн.

Деривациялық сипаттама қалмақ және орыс корпустарында шеттілдік терминдермен берілген: «Сөздің лексика-семантикалық сипаты сөзжасамдық белгіленіммен тығыз байланысты, бірақ басты мақсаты болып саналмайды. Сол себепті тек ең басты және жиі кездесетін деривациялық сипаттамалар берілді. Егер сөз «–го»-ға аяқталса, онда аталған лексикалық бірлік каритив деп белгіленді» [12] - деген құрастырушылардың пікірінен деривациялық белгілеуді сөз семантикасы үшін аса маңызды деп санамайтындықтары байқалады. Осыны ескере келе, орыс және қалмақ корпустарындағы деривациялық сипаттаманы беру үлгісі Қазақ корпусында басқаша көрініс таппақ.

Талқылау

Әлемдік корпустардағы лексика-семантикалық белігенімдерді талдай келе семантикалық және лексика-грамматикалық белгілеудің кең таралғанына көз жеткізуге болады. Бұл аталған сипаттамалардың лексика-семантикалық белгіленімнің өзегі болуымен байланысты. Қазақ

тілінің Ұлттық корпусында да осы дәстүр сақталады. Сонымен қатар Қалмақ тілінің Ұлттық корпусындағы аталған сөздердің коннотациясына қатысты белгілеу түрі қазақ тіліне де аса қажет деп танылды. Себебі қазақ тіліндегі етістіктерде жеке тұрғанда көрінетін эмоциялық реңк бар, ол сөйлем және мәтін мазмұнына тікелей әсер етеді. Шеттілдердің ұлттық корпустарындағы белгіленімді саралау арқылы қазақ тіліндегі етістіктердің лексика-семантикалық жақтан толық түсіндіре алатын белгілеу жүйесі әзірленді әрі қазақ тіліндегі белгіленім жасаудың негізгі шарттары анықталды. Қазақ тіліндегі етістіктердің лексика-семантикалық белгіленімі: тілдегі етістіктердің жалпы және жеке семантикасын анықтайды; лексика-грамматикалық сипаттама береді; коннотациялық реңкі анықталады; қазақ тілінің табиғатына сай семантикалық топтарға жіктеледі; кіші семантикалық топтар қолдануға ыңғайлы ірі класстарға біріктіріледі; корпустағы белгіленімдерді кодтауда көпшілікке түсінікті терминдер мен атаулар, ұғымдар қолданылады; етістіктердің омонимділігі ескеріліп, әр мағынасына жеке белгіленім жасалады.

Жоғарыдағы шарттарға сәйкес Қазақ тілінің Ұлттық корпусында лексика-семантикалық белгіленім төмендегідей 5 бағыт бойынша әзірленеді:

1.Сөзжасамдық сипатына қарай: дара/күрделі және негізгі/туынды. Сөз құрамы мен тұлғасын көрсету сөз мағынасын жақсырақ тануға мүмкіндік беретіндіктен белгіленім құрамына қосылып отыр. Нақтырақ айтқанда, берілген етістіктің түбір етістік немесе басқа сөз негізінде туындаған туынды етістік немесе бірнеше сөзден құралған күрделі етістік екендігін білдіретін грамматикалық сипаты негізінде туынды мағына анықталады. Мысалы: *«Ызыңда»* етістігі *«ызың»* еліктеуішінен жасалған туынды құрылым, сондықтан оның негізгі мағынасы – табиғаттағы жәндіктер шығаратын дыбысты білдіруі. Осы мағына негізінде оның лексика-семантикалық тобын айқындауға болады.

Ызыңда	етістік	*ара*	*туынды* *ызың+да*

2-кесте. Сөзжасамдық сипаттама

Сөзжасамдық белгі-код корпуста дәстүрлі терминдермен аталады, атап айтқанда *негізгі, туынды* және *дара, күрделі* болып түсінікті түрде жіктелінеді. Бұл мамандардың жаңа терминдерді меңгермеуінен немесе артқа тартуынан емес, керісінше, қолданушыға бұрыннан таныс ұғымдарды беру арқылы қолдануды барынша жеңілдету үшін жасалынып отыр. Себебі «аугментатив», «аттенуатив» секілді ағылшын терминдерін лингвистер түсінгенімен, басқа саланың мамандары түсінбеуі мүмкін.

2. Лексика-грамматикалық мағынасына қарай: болымды/болымсыз, салт/сабақты. Бұлай беру аталған категориялардың әрі семантикалық, әрі грамматикалық сипатқа ие лексика-грамматикалық категориялар болуымен тығыз байланысты. Яғни, олар сөзге тек грамматикалық қана емес, семантикалық мән үстейді, мағынасына жанама әсер етеді. Ғалым

М.Оразов та «Етістік» атты еңбегінде салттылық, сабақтылық категориясын семантикалық категория деп анықтайды [13, 158 б.].

3. Жеке (кіші) семантикасына қарай әр сөздің түпкі айырушы семасы анықталады, мағына терең танылады. Солардың негізінде 72 кіші семантикалық топтан тұратын классификация жасалды.

4. Жалпы (ірі) семантикасына қарай бөлу – кіші семантикалық топтарды ортақ семалары негізінде ірі класстарға біріктіру. Бұл іздеу процесін жеңілдетуге бағытталған. Орыс корпусының тәжірибесіне сәйкес, корпуста 5-6 ірі топтың берілуі қолданушының визуалды қабылдауына ыңғайлы. Қазақ тілі корпусында ондай ірі кластар саны – бесеу.

5. Коннотативті сипатына байланысты «жағымды», «жағымсыз», «бейтарап» деген үш белгі қойылады. Әр етістікте белгілі бір эмоциялық реңк болады және сол реңк оның мағынасына, контексте қолданылуына, синтактикасына тікелей әсер етеді. Кейбір синоним сөздер тек эмоциялық реңкі бойынша ғана ажыратылады. Ал кей етістіктер әр контексте әртүрлі реңк алуы мүмкін. Бұл негативті және позитивті мәнді ажыратуға ғана емес, тілімізде екі мәнде де қолданыла алатын етістіктерді анықтап көрсету арқылы қазақ тілінің мүмкіндіктерін таныту үшін қажет.

Лексика-семантикалық белгіленімнің ең маңызды бөлігі – сөз таптарын іштей семантикалық топтарға және субтоптарға жіктеу. Бұл процесс әр сөздің ішкі мағынасын тану негізінде жүзеге асады. Лексика-семантикалық белгіленім негізгі сөз таптарын қамтиды. Ғалым С.Исаев атап өткендей, «Тіл білімінің қай саласын алсақ та, оның зерттеу объектісі я тікелей, я жанама түрде сөзге келіп тіреледі. Өйткені сөзде фонетикалық қасиет (дыбыстық кешен) те, лексика-семантикалық қасиет (ішкі мағына) те, сөзжасам қасиеті (жаңа сөз я жаңа мағына) де, грамматикалық қасиет (морфологиялық құрамы мен құрылысы, сөздердің түрленуі, сол арқылы бір-бірімен байланысқа түсуі, алуан түрлі грамматикалық мағыналар білдіріп, сөйлем құрауы) те бір-бірімен қабысып, қабаттасып келіп отырады» [10, 5 б.]. Ғалым лексика-семантикалық жағын ішкі мағына деп атайды. Яғни лексика-семантикалық топтарға жіктеу тікелей бір сөзді екінші бір сөзден ажыратын дифференциалды семасын анықтай отырып топтастыруды білдіреді.

С.Исаев семантиканы лексикалық семантика және грамматикалық семантика деп бөліп қарастырады. Нақтырақ айтқанда, лексикалық мағынаға негізделсе, лексикалық семантика, ал грамматикалық мағынаға негізделсе, грамматикалық семантика болатындығына назар аударады. Корпуста – лексикалық семантика, яғни сөздің ішкі мағынасы назарға алынады. *Лексикалық мағына (семантика) ...сөздің ұғымдық мағынасы, сөздік мағынасы, сондықтан жалқы болады* [10, 287 б.].

Етістік – есімдерден кейінгі екінші орында тұратын грамматикалық үлкен категория. Түркі тілдеріндегі есім-етістік омонимиясын зерттеуші түркологтар түркі тіліндегі кез келген түпкі түбір әрі затты, әрі қимылды атап, синкретті болып келетінін айтады. Сондықтан да белгілі түрколог

А.М.Щербак етістік формаларының бастауында қимыл атаулары жатады дейді. Нақ осы қимыл атаулары есім-етістік синкреттілігі болған уақытында құбылыс пен процестерді тілдік құралдар арқылы атау барысында нақты сөйлеу үстінде сөйлеушінің іс-әрекетті орындаушыға және іс-әрекетке қарым-қатынасын білдірудің негізі қалана бастаған. Осылайша қимыл атаулары арқылы етістік формаларының семантикалық көп түрлілігі қалыптасқан [14, 96 б.]. Қазақ лингвистері етістіктерге ерекше назар аударып, көлемді зерттеулер жүргізген. Қазақ тіл білімінде етістік дәстүрлі принциптер тұрғысынан құрылымдық грамматикада және аспектуалдық тұрғысынан функционалды грамматикада қарастырылуда. Ал корпустық лингвистикада етістіктердің әрі грамматикалық, әрі лексика-семантикалық қыры ескеріліп, классификацияланады.

Етістіктердің лексика-семантикалық белгіленімін жасау процессі 7 кезеңнен тұрады. Олар:

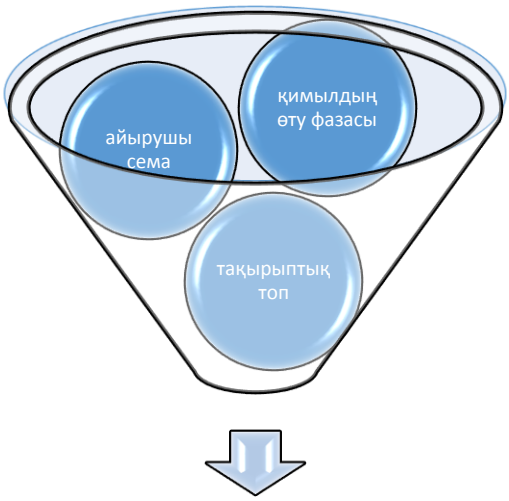
- бірінші: қазақ тіліндегі дара етістіктерді жинастыру;
- екінші: басқа тілдер корпусындағы лексика-семантикалық белгіленім жүйесін, үлгілерін талдау, сараптау;
- үшінші: қазақ және шетел тіл біліміндегі етістіктің лексика-семантикалық классификацияларын жинақтап, саралау, соның негізінде нақты мағыналық топтарды анықтау;
- төртінші: жинақталған етістіктердің лексика-грамматикалық сипаттамасын беру;
- бесінші: етістіктерді лексика-семантикалық топтарға жіктеу;
- алтыншы: жіктеуде қиындық тудырған етістіктер бойынша талдау жасау;
- жетінші: белгіленген алгоритм бойынша дайын мәліметтерді базаға енгізу, тексеру.

Осы уақытқа дейін бірқатар грамматистер етістіктерді мағыналарына қарай классификациялаған болатын. Мысалы, Ы.Е.Маманов етістіктерді сегіз топқа бөліп көрсетеді: 1) іс-әрекеттік қимылды білдіретін етістіктер (жазу, сызу, көру, жұлу т.б.); 2) қозғалыс-қимылды білдіретін етістіктер (жүгіру, еңбектеу, домалау); 3) қалып-күй процестерін білдіретін етістіктер (ауру, ұйықтау, отыру); 4) көңіл-күйді білдіретін етістіктер (қайғыру, мұңаю, қуану); 5) сапалық белгінің өзгерісін білдіретін етістіктер (ұзару, қысқару); 6) еліктеу-бейнелеу етістіктері (жылмыңдау, қылмыңдау, пысылдау); 7) туу, өсу қимылын білдіретін етістіктер (балалау, құлындау, боталау); 8) субъективтік реңдік етістіктер (азсыну, кемсіну, көпсіну) [15, 48 б.].

Берілген классификацияға сөйлеу, ойлау етістіктері енбеген. Сондай-ақ объектіге қатысты орындалатын әрекеттерден гөрі, субъектінің тікелей жасайтын әрекеттері назарға алынған. Сондықтан бұлай бөлу тілдегі барлық етістіктерді толық қамти алмайды. Ал ғалым А.Ысқақов «Қазіргі қазақ тілі» оқулығында «Етістіктің лексика-семантикалық сипаты» деген тақырыпта етістіктерді 11 мағыналық топқа бөледі. М.Оразовтың

жіктемесінде де етістіктердің 11 мағыналық тобы анықталады. Алайда белгіленім жасау мақсатында 18 200 етістік жинақталғанын ескере келе, оларды тек сегіз немесе он бір семантикалық топқа бөлгенде мағыналары толық ашылмайтындығы байқалды. Сол үшін лексика-семантикалық белгіленімді жасаушы топтың жетекшісі проф. Б.Момынова бұл мәселеге тереңдеп, топтардың санын 72-ге жеткізді. Оларды семантикалық топтар немесе субтоптар деп атауға болады. Бұл классификацияның басты артықшылығы әр етістіктің мәнін нақты әрі толық танытатын семантикалық тобы болуында. Кіші семантикалық топтарды Б.Момынова 5 ірі классқа (лексика-семантикалық топқа) біріктірді. Олар: *1) Адамның жай-күйін, қалыбын білдіретін етістіктер*, *2) Сөйлеу етістіктері*, *3) Көңіл-күй етістіктері*, *4)Қоғами-ғылыми етістіктер*, *5) Табиғатқа қатысты әрекеттер.*

Класстарға біріктірудегі негізгі мақсат – топтарды визуалды түрде қолдануға, іздеуге және түсінуге ыңғайлы түрде беру. Осы класстардың ішіне кіші 72 топ кіреді. Мысалы, «Табиғатқа қатысты әрекеттер» лексика-семантикалық тобы төмендегідей семантикалық топтардан және тілдік бірліктерден құралады: *1. Өсімдікке, егіндікке байланысты орындалатын амал түрлері* (егу, шабу, өсіру, шөптеу, суландыру, майқандау, малалау, миюлату, отау, себу, септіру, т.б.); *2. Үй жануарлары мен аңдардың жүрісі, қимылы* (шоқырақтау, жортақтау, лоқылдау, көткеншектеу, саржелу, жорғалау (бауырымен), азулау, тістелеу, қандау, тырналау, қарқырату, дал-далын шығару, үзу (сіңірін), шөгу, қорқырату, үйелеу, атылу, шиырлау, қорқаулану, лыпылдау, майпаңдау, мәлкілдеу, мөңку, мүйіздесу, сару, сауырлау т.б.**).** *3. Мал бағу, күтуге байланысты қимыл-әрекеттер* (жүгендеу, арқандау, шідерлеу, жүгіндіру, білдіру (бас), шалмалау, күзеу, күлтелеу, кісендеу, төлдету, қожырау, қайыру, шайт-шайттау, құраулау, қоректендіру, жаю, қылшықтау, құдилау, міну, қонжию, қотандау, қылтанақтау, қозықұлақтану, майсыздандыру (сүтті), матау, тағалау, өмілдіріктету, өру, т.б.). *4. Аңдар,үй жануарлары шығаратын дыбыстар* (қышқыру, кісінеу, мөңіреу, маңырау, маңырасу, бақылдау, шіңгірлеу, пысылдау, ырылдау, арқырау, шәңкілдеу, мәңкілдеу, мәуілдеу, осқыру, т.б.**);** *5. Үй және жабайы құстар, жәндіктер шығаратын дыбыстар мен оларға қатысты орындалатын әрекеттер* (ақтару (қоясын), үйрету, шашу (жем), бастырту (ұя), ұшу, қону, сусылдау, ауыздандыру, ысылдау, жыбырлау, жиырылу, ызыңдау, саяқсу, сүрлеу, тамақтау, т.б.); *6. Саятшылыққа байланысты әрекеттер* (салу (бүркіт), үркіту, қуалау, түсу (ізге), аңду, соғу (қамшымен) т.б.); *7. Табиғи апаттар мен құбылыстарға байланысты орындалатын қимыл-әрекеттер* (малмандай болу, малшыну, мәнерлеу, партылдау, себезгілеу, сел жүру, тасқындау, сөну, т.б.); *8. Табиғат сыйын пайдалану* (шайқау (май), консервілеу, масақтау, теру, жинау, мәйектену, самалдау, сәскелету, шомылу, т.б.); *9. Табиғаттағы аңдар мен құбылыстарға ұқсас қимылдар жасау* (қорбию, тарбию, пашырлату, рауандау, сақырлау, сарқылдау, сидаңдау, сойдию, судырау, сумаңдау, сырию, т.б.)



3-сурет. Семантикалық топтың негізі

Қорытынды

Лексика-семантикалық белгіленімнің әзірлемесін дайындауда практикалық жағына қатысты бірқатар мәселер туындады. Атап айтқанда, омоним және көп мағыналы етістіктерді контексте ажырату, бейнелеуіш мәндегі етістіктерді сипаттау, түрлі мағыналық реңкке ие етістіктерді нақты бір семантикалық топтың құрамына қосу және т.б. Аталған белгіленім Қазақ тілінің корпусында алғаш рет әзірленіп жатқандықтан мұндай күрделі тұстарының болуы заңды. Алғашқы кезеңде мақалада көрсетілген базалық сипаттамаларды беру белсенді түрде жүзеге асырылады. Алдағы уақытта жоғарыда санамаланған мәселелердің оңтайлы шешімі ретінде семантика-синтаксистік модельдер жасау қолға алынады.

Әлемдік корпус жасау тәжірибесі көрсеткендей лексика-семантикалық белгіленім – жасау өте күрделі және мәтінді автоматты талдау үшін аса маңызды белгіленім. Етістіктердің лексика-семантикалық белгіленімін әзірлеу қолданушыларға мынандай мүмкіндіктерді ашады:

– *Сөз мағынасын тереңірек тану;*
– *Кез келген уақытта электронды түрде қолдану;*
– *Синонимдес сөздерді анықтау;*
– *Тіл үйрену;*
– *Лексикографиялық еңбектер құрастыру;*
– *Семантикалық зерттеу жүргізу және т.б.*

Лексика-семантикалық белгіленім болашақта мәтін талдау процесінде маңызды рөлге ие семантикалық-синтаксистік модельдер жасаудың алғашқы қадамы болмақ. Параллель электронды ресурстарда семантикалық анализатордың болуы тілді әр қырынан қарастыруға, қазақ тілінің құрылысын, жүйесін жан-жақты зерттеуге жол ашады.

Зерттеушілер мен тілші мамандардың қажырлы еңбегінің арқасында күннен күнге дамып, қоры байып келе жатқан Қазақ тілінің Ұлттық корпусы тілдің мүмкіндігін танытатын, оны түсіну мен қолдануды жеңілдететін және зерттеуді жеделдететін қазақ тілінің басты цифрлы базасына айналуда.

ӘДЕБИЕТ

[1] Finegan E. Language: its structure and use. N. Y.: Harcourt Brace College Publishers, 2004. 611p.

[2] McEnery T., Wilson A. Corpus Linguistics. – Edinburgh, 2001.235 p.

[3] Захаров В.П. Корпусная лингвистика: Учебно-метод. пособие. – СПб., 2005. –48 с.

[4] Баранов А.Н. Корпусная лингвистика. Введение в прикладную лингвистику: Учебное пособие. – М., 2003. – 114 с.

[5] Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: Учеб. Пособие. – М., 2004. – 208 с.

[6] Сиразитдинов З.А. Моделирование грамматики башкирского языка. – Уфа, 2006. – 160 с.

[7] Meyer Ch. F. English Corpus Linguistics& An Introduction. Cambridge, 2002. – p.168.

[8] Kyto M., Rissanen M. A Language in transition: the Helsinki Corpus of English texts, ICAME Journal, 1992. Pp.7-27.

[9] Жұбанов А.Қ, Жаңабекова А.Ә. Корпустық лингвистика. – Алматы, 2017. – 336 б.

[10] Исаев С. Қазіргі қазақ тіліндегі сөздердің грамматикалық сипаты. – Алматы, 1998. – 304 б.

[11] Кустова Г. И., Ляшевская О. Н., Падучева Е. В., Рахилина Е. В. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. – М., 2005. – С. 155—174

[12] Куканова В. Принципы семантической разметки Национального корпуса Калмыцкого языка // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. – М., 2005. – С. 187—192.

[13] Оразов М. Қазақ тілінің семантикасы. – Алматы, 1991. – 216 б.

[14] Момынова Б. Қазақ тілінің морфологиясы. – Алматы, 2012. – 239 б.

[15] Маманов Ы.Е. Қазіргі қазақ тілі. – Алматы, 1966. 160 б.

REFERENCES

[1] Finegan E. Language: its structure and use. N. Y.: Harcourt Brace College Publishers, 2004. 611p.

[2] McEnery T., Wilson A. Corpus Linguistics. Edinburgh, 2001.235p.

[3] Zakharov V.P. Korpusnaia lingvistika. [Corpus linguistics]. Uchebno-metod. posobie. – SPb., 2005. 48 p. [In Rus.]

[4] Baranov A.N. Korpusnaia lingvistika. Vvdenie v prikladnuiu lingvistiku [Corpus linguistics. Introduction to applied linguistics]. M., 2003. 114 p. [In Rus.]

[5] Zubov A.V., Zubova I.I. (2004) Informacionnye tehnologii v lingvistike [Information technologies in linguistics] .Ucheb. Posobie. M., 2004. 208 p. [In Rus.]

[6] Sirazitdinov Z.A. Modelirovanie grammatiki bashkirskogo yazyka [Modeling of the grammar of the Bashkir language]. Ufa, 2006,160 p. [In Rus.]

[7] Meyer Ch. F. English Corpus Linguistics& An Introduction. Cambridge,Cambridge University Press, 2002. P.168.

[8] Kyto M., Rissanen M. A Language in transition: the Helsinki Corpus of English texts, ICAME Journal, 1992. Pp. 7-27

[9] Zhubanov A. K., Zhanabekova A. Korpustyq lingvistika [Corpus Linguistics]. Almaty, 2017. 336 P. [In Kaz.]

[10] Isaev S. Qazirgi qazaq tilindegi sozderdin grammatikalyq sipaty [Grammatical nature of words in the modern Kazakh language]. Almaty, 1998. 304 P. [In Kaz.]

[11] Kustova G.I., Lyashevskaya O.N., Paducheva E.V., Rakhilina E. V. Semanticheskaia razmetka leksiki v Nacional'nom korpuse russkogo yazyka: principy, problemy, perspektivy [Semantic markup of vocabulary in the National corpus of the Russian language: principles, problems, prospects]. National Corpus of the Russian language: 2003-2005 Results and prospects. M.,2005. PP. 155-174 [In Rus.]

[12] Kukanova V. Principy semanticheskoi razmetki Nacional'nogo korpusa Kalmyskogo yazyka [Principles of semantic markup of the National corpus of the Kalmyk language]. National Corpus of the Russian language: 2003-2005. Results and prospects. M.,2005. PP. 187-192. [In Rus.]

[13] Orazov M. Qazaq tilinin semantikasy [Semantics of the Kazakh language]. Almaty, 1991. 216 p. [In Kaz.]

[14] Momynova B. Қазақ tilinin morfologiyasy [Morphology of the Kazakh language]. Almaty: Arys publishing house, 2012. 239 p. [In Kaz.]

[15] Mamanov Y.E. Qazırgı qazaq tılı [Modern Kazakh language]. Almaty, 1966. 160 p. [In Kaz.]

РАЗРАБОТКА ЛЕКСИКО-СЕМАНТИЧЕСКОЙ РАЗМЕТКИ ГЛАГОЛОВ В НАЦИОНАЛЬНОМ КОРПУСЕ КАЗАХСКОГО ЯЗЫКА: МИРОВОЙ ОПЫТ, КЛАССИФИКАЦИЯ, РАЗМЕТКА В КОРПУСЕ

*Момынова Б.К.,[1] Имангазина М.А.,[2] Анесова У.Г.[3]

*[1]д.фил.н., профессор, Институт языкознания им. А.Байтурсынулы, Алматы, Казахстан,

[2]докторант 1-курса, Институт языкознания им. А.Байтурсынулы, Алматы, Казахстан,

[3]PhD доктор, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,

*[1]e-mail: momynova_b@mail.ru,

[2]e-mail: meruyertimangazina@mail.ru,

[3]e-mail: a_ymit@mail.ru

Аннотация. В статье рассматривается проблема разработки лексико-семантической разметки, одной из основных разметок в мировой практике построения корпуса. В частности, проводится обзор трудов отечественных и зарубежных ученых, касающихся компьютерной лингвистики и лексико-семантической классификации, показаны этапы создания лексико-семантической разметки глаголов в Национальном корпусе казахского языка, разъяснена практическая основа.

Ускоренные темпы развития информационных технологий требуют овладения электронными ресурсами и лингвистами. Областью лингвистики, изучающей и реализующей языковое программирование, является корпусная лингвистика. Создание Национального корпуса казахского языка основывается на создании разметок, которые автоматически анализируются по каждому уровню языка. Одной из сложных разметок в лингвистическом аннотировании слов является лексико-семантическая разметка. По сравнению с корпусом русского, калмыцкого и других

языков, лексико-семантическая разметка в Национальном корпусе казахского языка углубляется в значение слова, т. е. в сему. Поэтому количество малых (индивидуальных) лексико-семантических групп глагола составило 72 группы. Это позволяет пользователю более точно найти нужную ему информацию. Интерфейс применения системы разметки должен быть легок и понятен любому пользователю, как специалисту, так и специалистам других областей, которые только учатся пользоваться. Соответственно, лексико-семантические группы даются короткими и конкретными названиями.

В базу корпуса включены 18 200 глаголов, изучаются их смысловые оттенки. В ходе исследования было предложено дать характеристику глаголов по пяти различным признакам в лексико-семантической разметке. По словообразовательному характеру: простой, сложный; основной, производный; на основе лексико-грамматических категорий: переходность, непереходность; положительная и отрицательная форма; по характеру коннотации классифицируется как положительный, отрицательный, нейтральный. Для более глубокого раскрытия значения глаголов в зависимости от общих и отличительных сем они внутренне делятся на большие (лексико-семантические) и малые (семантические) группы.

Статья написана в рамках исследовательского проекта BR 11765619 «Разработка Национального корпуса казахского языка как информационно-инновационной базы государственного языка: научно-исследовательский и обучающий интернет-ресурс».

Ключевые слова: корпусная лингвистика, разметка, лексико-семантическая классификация, глагол, семантика, смысл, категория, лингвистическая аннотация.

DEVELOPMENT OF LEXICA-SEMANTIC MARKING OF VERBS IN THE NATIONAL CORPUS OF THE KAZAKH LANGUAGE: WORLD EXPERIENCE, CLASSIFICATION, MARKUP IN THE CORPUS

*Momynova B.K.,[1] Imangazina M.A.,[2] Anesova U.G.[3]

[1]Doctor of Philology, professor, The Institute of Linguistics named after A. Baitursynuly,Almaty, Kazakhstan,

[2]1st year doctoral student, The Institute of Linguistics named after A. Baitursynuly, Almaty, Kazakhstan,

[3]PhD doctor, Al-Farabi Kazakh National University, Almaty, Kazakhstan,

*[1]e-mail: momynova_b@mail.ru,

[2]e-mail: meruyertimangazina@mail.ru,

[3]e-mail: a_ymit@mail.ru

Abstract. The article deals with the problem of developing lexica-semantic markup, one of the main markups in the world practice of building a corpus. In particular, a review of the works of domestic and foreign scientists related to computational linguistics and lexica-semantic classification will be carried out, the stages of creating lexica-semantic markup of verbs in the National Corpus of the Kazakh language will be shown, and the practical basis will be explained.

The accelerated development of information technology requires the mastery of electronic resources in all branches of science, including linguistics. The corpus linguistics is the field of linguistics that studies and implements language programming. The creation of the National Corpus of the Kazakh language is based on the creation of markups, which automatically analyze each language level. One of the complex markups in linguistic annotation of words is lexical-semantic markup. Compared to the corpus of Russian, Kalmyk and other languages, the lexical-semantic markup in the National Corpus of the Kazakh

language deepens into the meaning of the word, i.e. into the sema. Therefore, the number of small (individual) lexica-semantic groups amounted to 72 groups. This allows the user to more accurately find the information he needs. The interface for using the markup system should be easy and understandable for any user, both a specialist in this field and specialists in other areas who are just learning to use it. Accordingly, lexica-semantic groups are given short and specific names.

The base of the corpus includes 18, 200 verbs, their semantic shades are being studied. In the course of the study, it was proposed to characterize verbs according to five different features in the lexica-semantic framework. First: by word-formation character, single, complex; main, derivative; the second: on the basis of the lexical and grammatical categories of transitivity, intransitivity; positive and negative form; connotative in character is classified as positive, negative, neutral. For a deeper disclosure of the meaning of verbs, depending on the common and distinctive semas, they are internally divided into large (lexica-semantic) and small (semantic) groups.

The article was written within the framework of the research project BR11765619 «Development of the National Corpus of the Kazakh language as information-innovation state language base: research and training internet resource».

Keywords: corpus linguistics, markup, lexica-semantic classification, verb, semantics, meaning, category, linguistic annotation.

Статья поступила 16.05.2022