ІСЛӘМ ЖАРЫЛҒАПОВТЫҢ 100 ЖЫЛДЫҚ МЕРЕЙТОЙЫНА АРНАЛҒАН
«ҚАЗАҚ СӨЗЖАСАМЫ: ТАРИХЫ, БҮГІНІ, КЕЛЕШЕГІ» АТТЫ ХАЛЫҚАРАЛЫҚ-ҒЫЛЫМИ ТӘЖІРИБЕЛІК КОНФЕРЕНЦИЯ МАТЕРИАЛДАРЫНЫҢ ЖИНАҒЫ, Қарағанды 2018 ж.

ҚАЗАҚ ТІЛІ МӘТІНДЕР КОРПУСЫНДАҒЫ СӨЗЖАСАМДЫҚ БЕЛГІЛЕНІМДЕР ҚОЮ БАҒДАРЛАМАСЫ

А.Ә.Жаңабекова
А.Байтұрсынұлы атындағы Тіл білімі институты,
Қолданбалы лингвистика бөлімінің меңгерушісі,
филология ғылымдарының докторы
Пірманова К.
А.Байтұрсынұлы атындағы
Тіл білімі институтының
2 курс магистранты

Қазақ тілінің ұлттық корпусын жасауда оның көлемін ұлғайту ғана емес, лингвистикалық мүмкіндіктерін кеңейту де маңызды мәселе. Лингвистикалық мүмкіндіктер деп отырғанымыз тіл деңгейлеріне сәйкес әртүрлі тілдік талдаулар түрін көбейту деген сөз. Ұлттық корпус дегеніміздің өзі тілдердегі мәтіндердің көлемді электронды жинағы ғана емес, сонымен қатар мәтінге автоматты түрде лингвистикалық талдаулар жасайтын компьютерлік бағдарламамен жұмыс істейтін ең алдымен лингвист-тілші мамандардың, білім алушылардың, тіл үйренушілердің кәсіби қажеттіліктерін өтейтін кең ауқымды ақпараттық-анықтамалық жүйе болып табылады. Мәтіндерге автоматты түрде белгіленім қоятын компьютерлік бағдарламаны лингвистикалық анализатор деп атайды. Тілдік талдаулар түрлеріне қарай морфологиялық анализатор, синтаксистік анализатор, фонетикалық анализатор, сөзжасамдық анализатор, лексика-семантикалық анализатор т.б. деп бөлінеді. Ал осы бағдарлама жасайтын лингвистикалық талдауларды «белгіленім» (разметка) деп атайды. Қай тілде болмасын алғашқы корпустарын құрастыруда ең алдымен морфологиялық анализатор бағдарламасы жасалып, морфологиялық белгіленімдер алдымен корпусқа енгізілген. Өйткені сөздің сөзтүрленімдік бөлігін әсіресе түркі тілдерінде бөлшектеу оңайырақ. Себебі түркі тілдері жалғамалы тілге жатады. Қосымшалар түбірден кейін ретімен бірінің үстіне бірі жалғанады. Мұндайда программа түбір сөздер мен оған жалғанған сөзтүрлендіруші қосымшаларды тез табады. Ал сөздің мұндай құрылымын бөлшектеу флективті тілдерді қиынырақ. Солай бола тұра, орыс тілінің ұлттық корпусында сөздерге морфологиялық белгіленімдер қоятын программа жасалған.

Қазіргі кезде А.Байтұрсынұлы атындағы Тіл білімі институты Қолданбалы лингвистика бөлімінде жасалған корпуста морфологиялық,

фонетикалық, лексика-семантикалық (біршама) және сөзжасамдық анализатор жасалды. Осы анализаторлар арқылы корпуста ізделген сөздің түбірі бөлініп көрсетіледі (лемматизация), сөзге жалғанған морфологиялық түрленімі бөлшектеліп, шартты белгілермен сипатталады; сөздер фонетикалық тұрғыдан да жуан және жіңішке әуезділігі, буын құрау бөліктері мен олардың түрлік сипаттамалары, жеке дыбыстардың талданымы көрсетіледі. Сонымен қатар туынды сөздердің негізгі түбірі мен сөзжасамдық бөлігін бөлшектей алатын әрі сөзжасамның қай тәсілі арқылы жасалып тұрғандығын сипаттайтын сөзжасамдық белгіленім бағдарламасы жұмыс істеп тұр. Мұндай лингвистикалық анализаторлар жұмысын қамтамасыз ету үшін программаға әр белгіленім түрі бойынша лингвистикалық нұсқаулықтар, яғни лингвистикалық базалар дайындау қажет. Морфологиялық анализатор бұдан бұрынғы жылдары жасалғандықтан, біз бұл жерде жаңадан жасалған белгіленім түрі – сөзжасамдық белгіленім енгізу бағдарламасы туралы баяндаймыз.

Қазақ тілінде түбір сөздер **негізгі түбір** және **туынды түбір** сөздер болып бөлінеді [1]. Мұның қай түрі де лексикалық бірлік ретінде сөздіктерге реестр ретінде енеді, мысалы, негізгі *мал* сөзі де, туынды *малшы* сөзі де дербес лексикалық бірлік, сондықтан екеуі де реестрлік тізімге кіреді. *Мал* сөзі сияқты қазақ тіліндегі негізгі түбір сөздерді сөзжасамдық анализатор реестр тізімінен іздейді. *Малшылық* деген туынды сөз кездескенде оны реестрдегі тізімге қарап отырып, оның *малшы* сөзінен туып отырғандығын, ал *малшы* сөзінің реестрде тағы кездесіп тұрған *мал* сөзінен туып отырғандығын көрсетеді. Яғни сөздің соңынан бастап, реестр тізімінде бар сөздермен салыстыра отырып, сөздің түпкі түбіріне дейін бөлшектеуі қажет. Демек, сөзжасамдық анализатор үшін де сол тілдегі түбір (реестр) сөздердің тізімін линвистикалық база ретінде беру қажет.

Қазақ тілінде сөздер жоғарыдағы *малшы* сөзіндегідей, тек қосымша жалғану арқылы ғана жасалмайды. Сондықтан сөзжасамдық белгіленім бағдарламасында, яғни сөзжасамдық анализаторға сөзжасамның басқа да тәсілдері туралы лингвистикалық нұсқаулық беру қажет.

Қазақ тілінде туынды сөздер үш түрлі тәсіл арқылы жасалды: *1) лексика-семантикалық тәсіл; 2) синтетикалық тәсіл; 3) аналитикалық тәсіл* [2].

Автоматты талдаулар үшін лексика-семантикалық тәсіл ескерілмейді, өйткені бұл тәсілмен жасалған сөздердің формальдық белгілері жоқ, сөз мағына кеңеюі арқылы жаңа мағыналарға ие болады. Корпус үшін сөзжасамдық белгіленім бағдарламасында синтетикалық және аналитикалық тәсіл арқылы жасалған туынды сөздер негізге алынады. Мұның ішінде синтетикалық тәсіл бойынша жасалған сөздерді талдауда программа реестр тізімі мен сөзтудырушы қосымшалар тізіміне сүйенсе, аналитикалық тәсілде көбінесе біріккен сөздер мен қос сөздердің тізіміне негізделеді, яғни программаға осындай сөздердің тізімі лингвистикалық база ретінде беріледі.

Алдымен сөзжасамның аналитикалық тәсіліне тоқталайық. Тілдегі аналитикалық тәсіл арқылы жасалған сөздерді күрделі сөздер дейміз. Мұндай

күрделі сөздер: *қос сөздер, біріккен сөздер, тіркесті түбір сөздер, қысқарған сөздер.*

Программа бұл сөздерді формальді белгілері арқылы да табуына болады. Мысалы, қос сөздер дефис арқылы жасалады. Қос сөздер қайталама және қосарлама қос сөз болып бөлінеді. Бұл екі түрі де дефис арқылы жасалғанмен, қайталама қос сөздер түбірдің қайталанып келуі арқылы жасалады, ал қосарлама қос сөздерде екінші сыңары басқа сөз болып келеді. Егер қос сөз құрамында екінші сыңарда бірінші сыңардағы сөзді қайталанып қолданса, программа оны қайталама қос сөз ретінде тануы қажет және соған сәйкес «қайталама қос сөз» деп көрсетеді. Мысалы: *қайта-қайта, әлсін-әлсін* т.б.

Белгіленім моделі: туынды, күрделі, анал., қос сөз, қайталама.

Ал егер қос сөздің екінші сыңары бірінші сынарынан мүлдем басқа сөз болса, яғни алдыңғы сыңарды қайталамаса, программа оны «қосарлама қос сөз» деп тануы қажет. Мысалы: *әке-шеше, ата-ана, бала-шаға* т.б.

Белгіленім моделі: туынды, күрделі, анал., қос сөз, қосарлама.

Біз жоғарыда қос сөздерді табудың бірінші жолын айтып отырмыз. Екінші жолы қайталама қос сөздер мен қосарлама қос сөздердің тізімін программаға лингвистикалық база ретінде беру. Мұндайда программа алдыңғы жолындай сөздердің екінші сыңарының қайталану/қайталанбай сипатына назар аудармайды, бұл жолы қос сөздің екі түрінің тізіміне ғана сүйеніп, мәтінде кездескен қос сөздерді ажыратады. Мұндай бағдарламада базаға берілген қос сөздердің тізіміне енбей қалған қос сөз белгісіз болып қалады, сөзжасадық белгіленім қойылмайды. Сондықтан қос сөздерді автоматты тануда осы аталған екі жолды да программаға енгізу қажет, яғни программа қос сөздерді осы екі жолмен де қатар іздеп көру керек.

Қос сөздерді формальді белгілері бойынша анықтағандай, компьютерлік бағдарлама біріккен сөздерді де тануына болар еді. Біріккен сөздер екі сөздің бірігуі арқылы жасалады. Рееестр сөздер тізіміне сүйене отырып, егер біріккен сөз құрамында екі дербес сөз қолданылып тұрса, мұндай сөздерді біріккен сөздер ретінде көрсетеді. Мәселен, саңырауқұлақ, итмұрын, аққала т.б.

Модель: туынды, күрделі, анал., біріккен сөз.

Ал кіріккен сөздерді бұл жолмен тану қиын, өйткені кіріккен сөздердің құрамы кірігіп, екі сөзге аражігі бөлінбейді, шегарасы көрінбейді, сондықтан кіріккен сөздерді программаға тек тізімдеп қана беру мүмкіндігі бар. Мысалы: *бүгін, бүрсігүні, әнеугүні т.б*. Қазақ тілінде мұндай кіріккен сөздер саны көп емес.

Модель: туынды, күрделі, анал., кіріккен сөз

Біріккен сөздерді автоматты танудың екінші жолы қос сөздердегідей олардың түрлеріне қатысты тізім жасау. Біріккен сөздерді кіріккен сөздерден басқа, өз ішінде жасалу тәсіліне қарап төл будан және кірме будан сөздер деп бөлуге болады. Төл будан сөздер дегеніміз – қазақ тілінің өз ішіндегі сөздердің бір-бірімен бірігіп, бір ұғымды білдіруі арқылы жасалған сөздер. Мысалы, аққала, ұзынқұлақ, көккұтан т.б. Кірме будан дегеніміз – біріккен

сөздің бір сыңарының басқа тілден енген сөз болып келуі арқылы жасалған сөздер. Мысалы: *авиакасса, автотұрақ, совхоз, партком т.б.*

Модель: туынды, күрделі, анал, төл будан/кірме будан.

Бұл аталған қос сөздер мен біріккен сөздерден басқа күрделі сөздердің ішіндегі қысқарған сөздер деген түрі де кездеседі. Қысқарған сөздерді танудың формальді белгісі бас әріптермен сөздің бастапқы әріптерінің жазылуы немесе нүктемен сөздің толық берілмеуі. Енді сөзжасамның синтетикалық тәсіліне тоқталайық. Тілдегі синтетикалық тәсіл арқылы жасалған сөздерді *дара туынды сөздер* дейміз. Синтетикалық тәсілде негізгі түбір сөздерге сөзжасамдық жұрнақтар жалғанып жаңа туынды сөз жасалады.

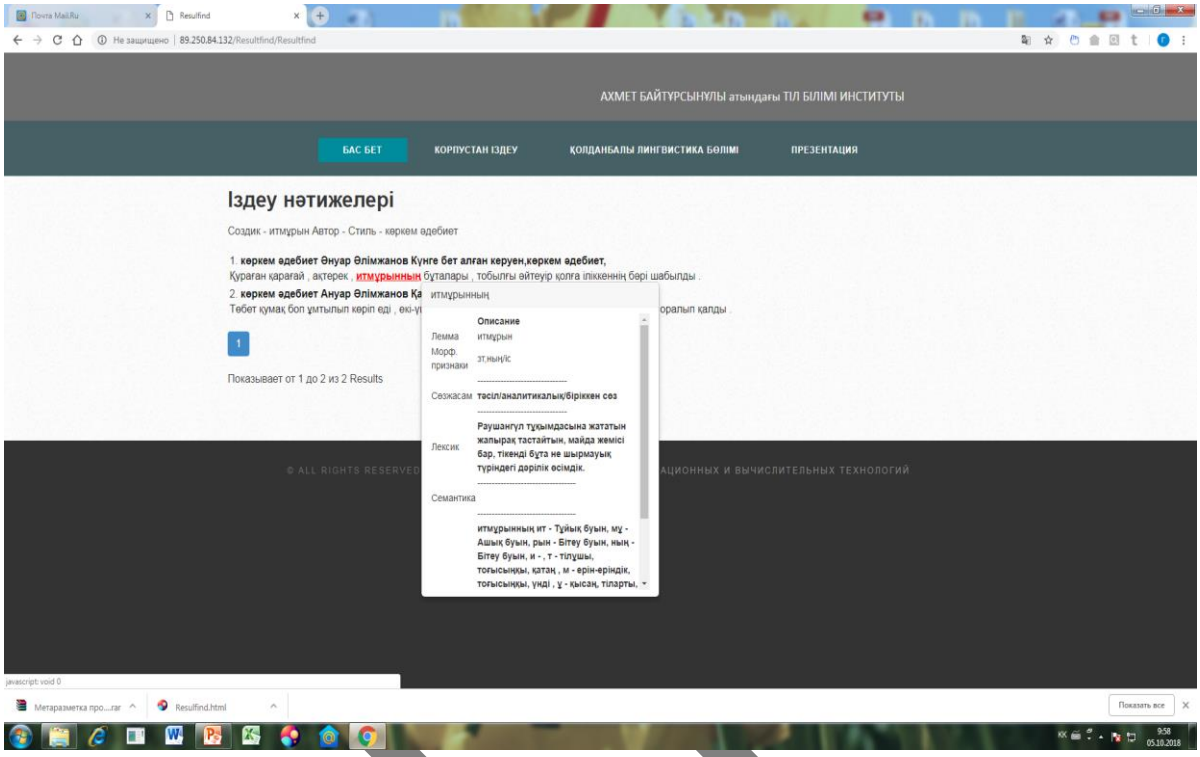
Бұл тәсілмен жасалған туынды сөздерді сөзжасамдық анализатор реестр тізіміндегі сөздерді және сөзжасамдық қосымшалардың тізімі мен сипаттамасы берілген кестені пайдаланып таниды. Біріккен сөздерді анықтауда программа сөз құрамындағы бөліктерді реестрден іздеп тапса, қосымша жалғану арқылы жасалған сөздерде де осы іздеу тәсілін қолдануға болады, яғни сөз құрамында бір ғана толық мағыналы сөз тұрса, толық мағыналы сөзден кейінгі жадғанып тұрған реестрде жоқ бөлік қосымша болып шығады. Мысалы: *малшы, малшылық, балалық, адамгершілік, абыройсыздық, балала, оюла, бірінші* т.б. сөздерде *мал, малшы, бала, адам, абырой, бала, ою, бір* сөздері реестрде бар толық мағыналы сөздер. Ал осы сөздерге жалғанып тұрған *-шы, -лық, -гершілік, -сыздық, -ла, - інші* бөліктері дербес сөздер емес, сондықтан программа оларды біріккен сөздер ретінде танымайды, екінші кезекте сөзжасамдық қосымшалар кестесінен осы қосымшаларды іздестіреді. Программа мұндай іздеуде әр сөз табының сөзжасамдық қосымшалар тізімі берілген кестеге сүйенеді. Қосымшалар кестесінен табылмаса, оларды тағы да қосымша бөліктеріне бөлшектеп көреді, мысалы: *-сыздық* қосымшасы кестеде жоқ, сол себепті анализатор оны *-сыз* және *-дық* деп тағы бөлшектеп, тағы да қосымшалар тізімінен іздестіреді. Әрі қарай іздестіру осылайша жалғаса береді.

Модель: **туынды, дара, синт. жұрнақ**

Сөзжасамдық жұрнақтарды әр сөз табына қатысты қысқаша шартты белгілермен беруге де болады. Мысалы: етістіктен етістік тудыратын жұрнақтар – **Ет.ет.туд.;** есімдерден зат есім тудырушы жұрнақтар – **Ес.зт.туд**. т.б.

Сөзжасамдық белгінім қоюда туынды сөздермен қатар негізгі сөздерді түпкі түбір екендігін көрсетуге болады. Негізгі түбірлер реестр тізімінде тұрған ешқандай қосымша жалғанбаған немесе ешқандай түбір сөз қосылып жазылмаған сөздің түпкі бөлігі. Мұндай негізгі түбірді – **Т** деп таңбалауға болады. Сөзжасамдық анализатор алдымен сөздердің осындай түпкі бөлігін іздеп тауып алуы қажет. Туынды сөз болмаса, ол негізгі түбірді бірден – **Т** (түбір) деп белгілеп, ал егер туынды сөз болса, оның негізгі түпкі түбірін бөліп алып, оған **Т** белгісін қойып, әрі қарай туынды сөзді жасаушы сөзжасамдық жұрнаққа қарай ығысады.

Енді біз төменде жоғарыдағы компьютерлік программаға берілетін лингвистикалық нұсқаулыққа сүйеніп жұмыс істейтін сөзжасамдық анализатордың нәтижесі – мәтіндер корпусындағы лингвистикалық терезе ұяшығында көрсетілетін сөзжасамдық белгіленімдерден кесінді береміз. ***1-сурет.***



Сурет 1 – Корпус мәтініндегі сөзге қойылған сөзжасамдық белгіленім

Суретте сөзжасамдық анализатор *итмұрын* деген сөздің біріккен сөз екенін, аналитикалық тәсіл арқылы жасалғандығын сипаттап көрсетіп тұр. Қорыта келгенде, корпус мәтігдеріндегі сөздерге автоматты түрде сөзжасамдық белгіленімдер қоятын программа жасау ұлттық корпус құрастыруда үлкен бір нәтижелі көрсеткіштердің біріне жатады. Бұл лингвистикалық зерттеулерде сөзжасам мәселелерін зерттеуде таптырмас ақпараттық-анықтамалық құрал болады. Бұл сөзжасамдық белгіленім қою программасын 100 пайыздық шындықты көрсетеді деп айта алмаймыз. Өйткені тілімізде әсіресе қосымшалар арқылы жасалған сөздерде омоним құбылысы өте жиі кездеседі. Сондықтан келешекте әркез түзету, жақсарту жұмыстарын үзбей жүргізіп отыру қажет.

Пайдаланылған әдебиеттер тізімі:

1 Қазақ грамматикасы. – Алматы, 2002.

2 Исаев С. Қазақ тіліндегі сөздердің грамматикалық сипаты. – Алматы, 1998.