ХҒТАР 16.31.31

Н. Абдикарим

Ш. Шаяхметов атындағы «Тіл-Қазына» ұлттық ғылыми-практикалық орталығы,
Астана, Қазақстан
E-mail: nursana@yandex.ru

Синтаксистік-семантикалық белгіленім жасаудың алғы шарттары

Аңдатпа. *Мақалада еліміздің цифрландыру саясатына орай әзірленіп жатқан қазақ тілінің корпустарын одан әрі жетілдіру, оны лингвистикалық зерттеулерде, оқыту үдерісінде және т.б. мақсатта пайдалану үшін халықаралық Лейпциг ережесіне негізделгенморфологиялықглостаудың шарттыбелгілеріниңтізімінжасап, ресмитүрде бекітудің қажеттілігі сөз болады. Қазақ тілі корпустарында және отандық компьютер лингвистикасы саласында пайдаланылып жүрген лингвистикалық метабелгілердің қолданысына талдау жасалады. Сонымен қатар төркіндес және т.б. тілдердің де осы бағыттағы еңбектері назарға ілініп, ондағы шартты белгілер мен белгіленімдердің сипаты мен ерекшелігі көрсетіледі. Морфологиялық глостауда қолданылатын шартты белгілер қазақ тілінің синтаксистік-семантикалық белгіленіммен жарақталған корпусын құруға негіз бола алады. Мұндай шартты белгілер ана тілімізді тірек тіл ретінде басқа тілдермен салғастыра немесе салыстыра зерттеуде тілдік материалдарды қысқа әрі көрнекілік нұсқада ұсынуға, грамматикалық белгілерін сипаттап жазуға, мәтіндік деректерді компьютерде өңдеуге; IT-технологияларды, әзірленген корпустарды пайдалана отырып лингвистикалық замануи зерттеулер жүргізуге жаңа мүмкіндіктер береді.*

Түйін сөздер: *цифрландыру саясаты, тіл корпусы, Лейпциг ережесі, глостау, морфологиялық тег, шартты белгілер, синтаксистік-семантикалық белгіленім, заманауи лингвистикалық зерттеулер.*

DOI: https://doi org/10.32523/2616-678X-2023-144-3-08-21

Кіріспе. Өткен ғасырдың 60-жылдарынан машинамен мәтін аударудан басталған үдеріс қазіргі уақытта кез келген тілді компьютерде автоматты түрде өңдеу, Ғаламтор желісінен ақпарат алу, смартфон пайдалану, жасанды интеллект көмегімен мәтін өңдеу, чатботтар арқылы сұхбаттасу, білім беру үдерісінде цифрлық технологияларды қолдану тәрізді қоғамның сұранысы мен қажеттілігін өтейтін өнімдерді ұсына алатын ақпараттық технологиялар дәуіріне ұласты. Табиғи тілді компьютерде пайдалану, өңдеу, сандық технологияларды қолданудың арқасында автоматтандырылған аударма, мәтіндік деректер базасын жасау, оны алуан түрлі мақсатта пайдалану мүмкіндігі ашылды. Осыған орай кез келген ғылым саласында дәстүрлі зерттеулердегі мәселелерді жаңа қырынан қарастыру қажеттілігі туындады.

Қазақ тіл білімінің алдына да төл тіліміздің Ұлттық корпусын жасақтау және оның белгіленім тереңдігін ұлғайту, ішкі және сыртқы белгілерін есепке алу, әр түрлі кіші корпустарын жасау тәрізді көптеген міндеттер қойылып отыр. Сонымен қатар қазақ сөзін дыбыстық және графикалық тану, омонимдерді ажырату, мәтінді талдау мен синтездеудің, семантикалық-синтаксистік талдағыштардың жетілдірілген әрі қазақ тіліне лайықталып бапталған, автоматтандырылған жүйелері мен ережелерін әзірлеу мәселесі күн тәртібінде тұрғаны белгілі.

Өткен ғасырдың 30-шы жылдарынан бастап Қ. Жұбанов, С.Аманжолов, Ж. Сәдуақасұлы, Б. Қайыров, А. Жұбанов және т.б. ғалымдарымыз қолданбалы лингвистиканың негізін қалап, сөйлемді формалдау мен модельдеудің әр түрлі үлгілері мен шартты белгілерін ұсынып, статистикалық әдістермен зерттеді. Соңғы кезде формалданған сөйлем үлгісін пайдаланып, мысалы, сөйлемді терең және үстірт құрылымға бөле отырып, субъект – N, предикат– V, анықтауыш – A, объект – O және т.б. шартты белгілермен көрсетіп, қазақ тілін өзгетілді аудиторияға арнайы алгоритм арқылы меңгерту жолын қарастырып жүрген зерттеушілер де бар екендігі жайында бұған дейінгі зерттеулерімізде де айтқан болатынбыз[1, 83-85]. Осы үрдісті әрі қарай жетілдіріп, тіліміздің бәсекеге қабілетті, заман талаптарына сай зерттеулер көшіне ілесе алатындай дәрежеге жеткізу үшін әрі еліміздегі цифрландыру саясаты мен мемлекеттік тілді дамытудың ұзақ жылдарға арналған бағдарламаларын жүзеге асыру үшін тіл корпустары жасалып жатыр. Соңғы кездері «... корпус технологиясының дамуына байланысты лингвистикалық зерттеулердің эмпирикалық және әдістемелік базасы кеңейіп, аталған сала үшін жаңа мүмкіндіктер ашылды» [2, 138]. Бұл мүмкіндіктер әзірленген корпусты пайдаланып, ғылыми зерттеу жұмыстарын жүргізу, статистикалық мәлімет алу, мәтінді талдау, өңдеу, сөйлемді немесе оның құрылымын формалдау, модельдеу, оқу үдерісіне пайдалану немесе оның бағдарламасын жасақтау сияқты көптеген сұрақтардың шешімін табуға көмектеседі. Осылардың барлығының өзегінде табиғи тілді шартты белгілердің көмегімен машина түсінетіндей етіп ұсыну мәселесі жатыр. Қазіргі қазақ тілінің корпустары грамматикалық леммалармен белгіленгенімен, ондағы тілдік бірліктерді халықаралық стандарттарға сәйкес толыққанды өрнектеу, яғни тегтеу тетіктерінің бірізге түспендігі біздің мақаламыздың жазылуына алғы шарт болып табылады.

Біз сөз етіп отырған морфологиялық глостаудың шартты белгілері жайындағы мәселе,біріншіден,екітүрлібілімсаласының – компьютер және лингвистиканың түйісуінен пайда болған, бірақ табиғи тілдің деректерін өз мақсаттарына сай пайдаланатын ғылымның екі саласының мамандарының (біздің еліміздің) әлі де болса бірлесіп жұмыс жасауға құлшынысы төмен екендігінен; екіншіден, осы бағытта жазылып, жарық көріп жатқан еңбектердегі шартты белгілер тіліміздің табиғатын формалды түрде толықтай өрнектеп шығуға, белгіленім жасауға (аннотациялауға) толық дайын еместігінен, яғни шартты белгілер жүйесі жасалып, бекітілмегендігінен туындап отыр.

Компьютер мен IT-технологиялар индустриясы қазақ лингвистикасында: 1) компьютер лингвистикасын дамытуды көздейтін IT-бағдарлама жасақтаушылар; 2) корпус әзірлеумен айналысып жүрген лингвист мамандардың арқасында екі бағытта дамып келеді және олардың бірлесіп атқарған жоба немесе жарық көрген елеулі еңбектері жоқтың қасы деуге болады. Осы олқылықтың орнын толтырып, тиісті сала мамандарының күшін біріктіру мақсатында ҚР тіл саясатын іске асырудың 2020-2025 жылдарға арналған бағдарламасында мынадай міндет қойылған: «... лингвистер IT-мамандармен бірлесіп морфологиялық, морфо-семантикалық, лексика-семантикалық, синтаксистік, фонетикалық, мәдени-семантикалық белгілер мен метабелгілер жүйесін әзірлейтін болады [3]. Демек, бұл мәселеде пәнаралық зерттеулер мен талқылаулардың, ортақ қағидалар мен ережелердің болатыны ләзім. Сондықтан да ең әуелі осыған дейін елімізде компьютерлік

лингвистикасымен шұғылданып жүрген ғалымдарымыз бен IT-мамандардың қазақ тілінің корпустарында пайдаланып жүрген шартты белгілеріне талдау жасап, зерделеп алған дұрыс.

Айталық, «Мемлекеттік тілдің ақпараттық-инновациялық базасы ретіндегі қазақ тілінің ұлттық корпусын әзірлеу: ғылыми-зерттеу және оқыту интернет-ресурсы» жобасы аясында жазылған «Ұлттық корпустарға негізделген лингвистикалық зерттеулер жүргізу» деп аталатын мақалада электрондық тілдік корпустарды лексикографиялық зерттеулердепайдаланумүмкіндіктері, яғни оны әртүрлі сөздіктер (терминологиялық, этнографиялық, аймақтық, түсіндірме т.б.) құрастыру тәжірибесінде пайдаланудың лингвистикалық және инженерлік технологиясы, әдіс-тәсілдері айқындалатындығы [4, 84-85] және т.б. маңызды мәселелер шешімін табатындығы сөз болады. Алайда мұнда лингвистикалық міндеттерді шешуден гөрі компьютер лингвистикасының ғылыми-инженерлік мәселесіне көбірек назар аударылған. Өйткені компьютер лингвистикасында тілдің инфлективтік және деривациялық көрсеткіштеріне аса көп көңіл бөлінеді де, тіл фактілерін шартты белгілермен тегтеуден гөрі негізгі тілдік бірліктерді, олардың бөлшектерін дайын қалпында программаға енгізу маңызды болып табылады. Негізінен арнайы компьютер бағдарламалары арқылы лингвистикалық міндеттерді автоматты түрде шешу үшін төмендегі 4 бағдарлама қолданылады екен: 1) іздеу үдерісін автоматтандыруға арналған сөздіктер мен тезаурустардың электрондық нұсқасы; 2) мәтіндерді түрлендіру немесе генерациялауға, аннотациялау мен рефераттауға арналған бағдармалар; 3) мәтінді өңдеу мен лингвистикалық талдауды (мәтіндегі орфографиялық, грамматикалық, стилистикалық қателерді тексеру, тасымалдау, қарапайым статистикалық талдау жасау, сөздерді лемматизациялау, синтаксистік және морфологиялық талдау) жүзеге асыратын бағдарлама; 4) мәтіннің мазмұнын түсінуге және оған нақты жауап беруге, яғни жасанды интеллектіні дамытуға бағытталған табиғи тілді өңдеу жүйесі [5, 40-41]. Осы мақсатта қазақ тіліндегі мәтіндердің лингвистикалық процессорын әзірлеу үшін бұлттық веб-сервизді IT-инструменті ретінде қолдану, яғни мәтіндік контенттердегі адамдардың сын-пікірін талдау, қоғамдық көңіл-күйін бағалау үшін мәлімет іздеу, прагматикалық қолданыс үшін контент-анализ жасау, таргет-жарнама, локализация т.б. мақсатында қазақ сөздерінің лексика-морфологиялық анализі жасалып жатыр. Дегенмен компьютер саласының мамандарының қазақ тілінің морфологиялық белгіленімі бойынша атқарып жатқан жұмыстарын жоққа шығаруға болмайды. Мысалы, 2013 жылы түркі тілдес елдердің компьютер лингвистикасы саласындағы ғалымдар бас қосатын «TurkLang» конференциясы ұйымдастырылып, келесі жылында түркі тілдеріндегі электрондық мәтіндер үшін бірізідендірілген грамматикалық белгіленім жүйесін әзірлеу туралы Қарар қабылданған. Сол кезде түркі сөзформаларының морфемдық құрамына негізделген морфологиялық белгіленімнің алғашқы нұсқасы таныстырылды және ол барлық түркі тілдерінің морфологиялық ерекшеліктерін көрсете алатындай болу керектігі келісілді [6, 80].

Ал отандық лингвист мамандар тілдік корпустарды жасақтау, мәліметтерді жинақтау, база құрастырумен айналысуда: А. Байтұрсынұлы атындағы Тіл білімі институты ана тіліміздің барлық стилі қамтылған, мәтін көлемі 21 млн. болатын Қазақ тілінің Ұлттық корпусын [7] әзірлесе, Ш.Шаяхметов атындағы «Тіл-Қазына» ұлттық ғылыми-практикалық орталығы «Қазақ тілінің ұлттық корпусының бес кіші корпусы» жобасын қолға алып, қазіргі кезде оның публицистикалық кіші корпусын әзірленіп жатыр және жобаның мәтіндік базасының көлемі 40 миллион сөзқолданысқа жеткізу межеленіп отыр [8].

IT-технологияларды таза лингвистикалық мақсатта қолдану үшін ең әуелі тілдің «... барлық инфлективтік және деривациялық көрсеткіштерін сипаттау және белгілеп алу; екіншіден, алломорфтарды таңдап алу ережесін, сөзформаларын автоматты талдауға

арналған сандхи[1] (морфема шекарасындағы морфонологиялық үдерістер, бір ғана сөзформасындағы фонетикалық үдерістер) ережелерін әзірлеу қажет [9, 7]. Жоғарыдағы «TurkLang» конференциясының Қарары аясында жүзеге асырылып жатқан жұмыстар бойынша елімізде жарық көрген бірлі-жарым мақалалар болмаса, жүйелі түрде жинақталған мәліметтерді кездестірмедік. Ал «Тюркская морфема» деп аталатын порталда негізінен татар тілінің деректері тіркелген. Осы айтылған мәселелерді саралай келгенде, қазақ тілінің морфологиялық белгіленім жүйесін одан әрі жетілдіріп, қолданыстағы белгіленімдерге талдау жасап, оны компьютер лингвистикасы ғана емес, лингвистикалық типологияның ғылыми-зерттеу жұмысына пайдаланудың маңыздылығын және ол синтаксистік-семантикалық белгіленім жасаудың алғы шарты болып табылатындығын көреміз.

Зерттеу әдістері. Қазақ тілінің белгіленім жасалған корпустарында және отандық компьютер лингвистикасы саласының мамандары қолданып жүрген шартты белгілер мен белгіленімдерге талдау жасалып, морфологиялық глостауды біріздендіру жолдарын қарастыру және оның лингвистикалық заманауи зерттеулердегі маңызын анықтау зерттеменің негізгі міндеттерінің бірі болып табылады. Осыған орай компьютерлік немесе сандық лингвистика, қолданбалы лингвистиканың әртүрлі салаларын түйістіретін жасанды әрі шартты метатілдік белгілерді жүйелеу әдістемесі негізге алынды. Оның ішінде корпустық лингвистиканың морфологиялық тегтеу әдісінде қолданылатын метабелгілерді синтаксистік және семантикалық шартты белгілер жасақтауға бейімдеу, оны қазақ тілінің табиғи заңдылықтарына сай таңдау жолдары көрсетілді. Сөз болып отырған мәселе қазақ тілін формалдау үдерісін халықаралық жасанды тілге сай оңтайландыру, тілдік құбылыстарды автоматты түрде өңдеу немесе модельдеуді нысанаға алып отырғандықтан, интра-лингвистикалық әдістемеге де негізделеді. Сонымен бірге тіл және оның белгілі бір аспектілері бойынша зерттеу жүргізуде кеңінен қолданылатын шолу, бақылау, жинақтау әдістері; метатілдік деректерді сипаттауға қажетті шартты белгілерді салыстыру, оларға талдау жасау үшін анализ, синтез және сөз болып отырған мәселеніжан-жақтысипаттау, ғалымдардың алуан түрлі пікірлерін жалпылау әдістері пайдаланылды.

Талдау. Мәтіндегі кез келген сөзге морфологиялық ақпаратты тіркеу төрт «алаң» немесе белгілер тобынан тұрады: лексеманың сөз табына қатыстылығы, грамматикалық белгілерінің жиынтығы, сөзтүрлендіруші сипаты, грамматикалық тұлғалардың стандарт, нормадан ауытқуын көрсететін белгілер [10]. Негізінен корпустарда пайдаланылып жүрген шартты белгілер кез келген тіл мамандырының қолданып, тілдік фактілерді «оқи алуына» жағдай жасайтындықтан, латын әліпбиі негізінде қысқатылған шартты белгілер – тегтер қолданылады; олар метатілдің негізін құрайды. Ал тегтердің өзі морфологиялық глостауды жүзеге асыратын Лейпциг ережесіне негізделеді.

«Глосс» деген түсінік алғашқыда мәтіндердегі ескірген, түсініксіз сөздерге (талданып отырған сөздің үстінде немесе астында орналасқан интерлинеарлық глосс және шетіндегі маргинальдық глосс) түсініктеме беру практикасының негізінде пайда болған. Қазіргі жағдайда глосс дегеніміз көбінесе типологиялық зерттеулерде бірнеше тілді салыстыра немесе салғастыра зерттеу барысында келтірілген тілдік мысалдардың құрамындағы сөздер мен морфемдердің тиісті грамматикалық мәліметін қоса беретін «аудармалар» деуге болады; ол мәліметтерталданыпотырғанморфеманың/сөздің грамматикалық (категориялдық) белгілерін көрсететін арнайы қабылданған символдар мен қысқартулардың көмегімен «шифрланады» [11, 219]. Кез келген ұлттық корпустарда глостаудың халықаралық Лейпциг ережесіне негізделген лексика-морфологиялық белгіленім жүйесі кеңінен қолданылады, онда типологиялық кең

[1]Сандхи – (санскрит IASTsaṃdhi sa. संधि қосылу) морфема немесе сөз шекарасында орын алатын фонологиялық процестердің кең ауқымын қамтитын термин.

таралған грамматикалық категориялар қысқартып белгіленеді және белгілі бір тілдің теггеу жүйесін сол Ережеге барынша жуықтастырып жасақтауға тырысады. Лейпциг ережесі (барлығы 11 ереже) тілдік бірліктерді сипаттау стандартын ғана белгілейтіндіктен оны әр тіл өзінің заңдылықтары негізінде өзгертіп, ыңғайлап қолдануға да болады. Демек қазақ лингвистикасы үшін мұндай «шифрлардың» қажеттілігі сөзсіз, яғни глостық символдар мен қысқартылған шартты белгілердің жүйесі мен тәртібін төл тіліміздің ерекшелігіне, грамматикалық белгілеріне сай таңдап, жүйелеп алуымыз керек.

Түркі тілдері морфологиясының біріккен метатілін әзірлеуді мақсат тұтқан ғалымдарымыз бен отандық тілдік корпус құрастырушалырдың пайдаланып жүрген шартты белгілер – теггерге талдау жасап көрейік.

1) А. Шарипбай бастаған ғалымдар тобының бір мақаласында түркі тілдерінің көптілді тезаурус жасауға қажетті теггері былайша берілген: қысқартылған тег, ағылшын, қазақ-татар-қырғыз-өзбек-түрік және орыс тіліндегі зат есім және оның құрамы бойынша атаулары.

2) А. Байтұрсынұлы атындағы Тіл білімі институты әзірлеген қазақ тілінің Ұлттық корпусында сөз таптары (10)[2], септік жалғау (7), сан (3), тәуелдік жалғау (4), жақ (3), шырай (2), шақ (3), рай (4), етістіктің болымсыз (1) формасы, тұйық етістік (1), есімше (1), көсемше (1), функционалды қосымшалардың (9) қазақ тіліндегі қысқартылған шартты белгілері, қай сөз табынан екендігі, орысша атаулары, қысқартылған халықаралық шартты белгісі көрсетілген.

3) Ш.Шаяхметоватындағы «Тіл-Қазына» ұлттық ғылыми-практикалық орталығының Қазақ тілінің публицистикалық кіші корпусында шартты белгілер айдарында сөз таптары, қазақша қысқартылған шартты белгілері, орысша атаулары, халықаралық стандарт бойынша белгіленуі (толық атауы) туралы мәлімет берілген. Тілдік бірліктердің қамтылу саны жағынан жоғарыда аталған Ұлттық корпусқа ұқсас болғанымен, тілдік деректердің біршама толыққанды сипатын беру көзделгендігі құптарлық. Мысалы, көмекші есімдер де сөз табы ретінде көрсетіліп, сын есімнің шырайлары *жай* және *асырмалы* шырайлармен толықтырылған. Сонымен қатар етіс формалары, сөзжасамдық жұрнақтар; одағай, етістік, сын есім, есімдік, үстеу, еліктеуіш, шылау, зат есім семантикасы және оның құрамына қарай жіктелген түрлері қосылған.

Осы еңбектердегі шартты белгілерді салыстыра отырып мынаны байқадық: 1-ші топта морфологиялық бірліктердің ағылшын тіліндегі атаулары мен қысқартылған нұсқасы қатар берілген; 2-ші топта халықаралық шартты белгілердің қысқартылған нұсқасы берілсе, соңғы топта керісінше ол нұсқасы көрсетілмеген. Сонымен қатар тілдік бірліктердің тізілімі мен саны да әр түрлі екендігін төмендегі кестеден көруге болады:

[2]Жақшаның ішінде грамматикалық категорияның саны көрсетілген (Сөз табы (10) - 10 сөз тобы және т.б. дегенді білдіреді).

Түркі тілдеріне ортақ метатіл			*Публицистикалық кіші корпус*		
атауы	*ағылшын*	*тег*	*атауы*	*ағылшын*	*шартты белгі*
Зат есім	Noun	**N**	Зат есім	Noun	зт.
Дара зат есім	Simple Noun	**SIMP**			-
Күрделі зат еіс	Complex Noun	**CMPL**			-
Біріккен зат есім	Fused words	**FUSW**	Біріккен	Closed compounds	-
Қосарланған зат есім	Pair Noun	**PAIR**	Қосарланған	Hyphenated compounds	-
Тіркесті (құрама) з.е.	Compound Noun	**CMPN**	-		-
Қысқарған з.е.	Abbreviations	**ABBR**	-		-
Негізгі з.е.	underivatives Noun	**UND R**	-		-
Туынды з.е.	derivatives Noun	**DRVT**	-		-
Жанды з.е.	animate Noun	**ANIM**	Адамзат есімдері	Animate nouns	Адз
Жансыз з.е.	inanimate Noun	**INAM**	Ғаламзат есімдері	Inanimate nouns	Ғалз
-	-	-	Деректі з.е.	Concrete nouns	Дер
-	-	-	Дерексіз з.е.	Abstract nouns	Дерз

Кесте 1. Зат есімнің метатілдік шартты белгілері

Болашақтағы типологиялық зерттеулер мен синтаксистік белгіленім үшін қажетті шарттыбелгілер дайындаудатүркітілдеріне ортақ метатіл әзірлеушілердің ұсынып отырған нұсқасы маңыздырақ екендігін атап өткен жөн. Ал публицистикалық кіші корпуста шартты белгілер қазақ тілінде қысқартылып, адамзат есімдері мен ғаламзат есімдері, деректі және дерексіз зат есімдердің ажыратылып берілгені тілдің семантикалық сипатын ажыратуға септігін тигізеді.

Жалпы «Компьютер коммуникациясының мәтіндері сигналдық қана емес, әр түрлі икондық (семиотикалық ұғымдағы –А.Н.)) белгілерді де қамтиды, соңғысы назар аударарлық, сәндік, нақтылау және символдық функцияларды орындайды» [12, 7]. Осылардың ішінен мәтіндегі грамматикалық тұлғаларды нақтылау функциясы лингвистикалық әрекеттер мен зерттеулерге қатысты деп түсінуге болады.

Біздің бірінші мақсатымыз халықаралық стандартқа негізделген әрі төл тіліміздің ерекшелігін көрсетіп, нақтылап бере алатындай және төркіндес түркі тілдері мамандары да қиындықсыз тани білетіндей шартты белгілердің жүйесін әзірлеп, оны біріздендіруге көңіл бөлу болып табылады. Ендеше әр түрлі тілдер корпусында қолданылып жүрген сөз таптарына қатысты тегтер мен Лейпциг ережесіндегі морфологиялық глостық белгілерді салыстырып көрейік:

№	Сөз таптары	Орыс тілінің Ұлттық корпусы	Татар тілі [13]	Саха тілі [9, 7]	Башқұрт тілі [14, 138-139]	Лейпциг ережесі [15; 16]
1	Зат есім	S	N	N	S	N
2	Сын есім	A	ADJ	ADJ	ADJ	ADJ
3	Сан есім	NUM	NUM	NUM	NUM	NUM
4	Етістік	V	V	V	V	V
5	Үстеу	ADV	ADV	ADV	-	ADV
6	Жалғаулық	CONJ	CNJ	CONJ	CONJ	CONJ/ CNJ
7	Демеулік	PART	PART	PART	PART	PART
8	Одағай	INTJ	INTRJ	INTRJ	INTJ	INTRJ
9	Есімдік	-	PN	PN	SPRO	PRO
10	Сан-сын есім	ANUM	-	-	-	-
11	Предикатив	PRAEDIC	-	-	-	PRED
12	Қыстырма сөз	PARENTH	-	-	-	-
13	Есімдік-зат есім	SPRO	-	-	-	-
14	Есімдік-сын есім	APRO	-	-	-	-
15	Есімдік-үстеу	ADVPRO	-	-	-	-
16	Есімдік – предикатив	PRAEDICPRO	-	-	-	-
17	Предлог	PR	-	-	-	-
18	Септеулік	-	POST	POST	POST	POST
19	Модаль сөз	-	MOD	MOD	-	MOD
20	Еліктеуіш сөз	-	IMIT	IMIT	-	IMIT
21	Есімше	**partcp**	-	PCP	-	PTCP/PCP
22	Көсемше	**ger**	-	CONV	-	CVB
23	Тұйық етістік	**inf**	-	-	-	INF
24	Поссесив (тәуелділік маркері)	-	-	POSS	POSS	POSS

Кесте 2. Сөз таптарының қысқартылған шартты белгілері

Жоғарыдағы кестеде орыс тілінің Ұлттық корпусында, татар тілі корпусында [13] және саха тілінде [9, 7], башқұрт тілінің морфологиялық анализаторында қабылданған [14, 138-139] тегтер сондай-ақ глостаудың Лейпциг ережесі [15; 16] бойынша сөз таптарын белгілеудің қысқартылған шартты белгілері [7, 8, 11] салыстырылды. Орыс тілінің корпусында 1-17 (есімдікті қоспағанда) аралығындағы шартты белгілер, ал татар және саха тілдерінде 1-9, 18-24 аралығынндағы грамматикалық категориялар сөз табы ретінде көрсетіліп, қысқартылған шартты белгілері қоса берілген; татар тілінде есімше, көсемше, тұйық етістік, поссесив тұлғалары сөз табы ретінде көрсетілмеген болса, саха тілінде инфинитив жеке сөз табы ретінде белгіленбеген; туыстас екі тілдің жалғаулықтарды, есімше мен көсемшенің шартты белгісі Лейпциг ережесіндегі екі нұсқаны бір-бірден қабылдағандығы байқалады; башқұрт тілінде зат есім мен есімдіктің шартты белгісі орыс тілінікімен бірдей әрі үстеудің тегтік белгісі жоқ, ал қалған шартты белгілер аталған Ережемен негізінен сәйкес келеді. Орыс тілінде сын есім немесе есімдіктің сөйлемдегі семантикалық мағынасынан туындайтын граммемалар біршама нақты көрсетілген және мұндай шартты белгілер дүниежүзілік стандарт болып табылатын Лейпциг ережесінде

жоқ. Және де есімше, көсемше, инфинитив тұлғалары етістіктің бір көрінісі ретінде кіші әріптермен қысқартылып берілген. Бұл орыс тілінің корпусы «Смысл-Текст» деп аталатын А. Мельчуктің әйгілі теориясын басшылыққа алатындығымен байланысты болса керек.

Замануи лингвистикалық әдебиеттерде кең тараған типологиялық грамматикалық категорияларды белгілеу үшін Бернард Комри, Мартин Хаспельмат және Бальтазар Бикелдер әзірлеген глостаудың Лейпциг ережесіндегі қысқартылған грамматикалық аббервиатуралар пайдаланылады және ол халықаралық стандарт болып табылады [17]. Корпустарда қолданылатын лексика-морфологиялық белгіленімдер барынша осы ережеге жуықтастырылып жасалуы керек. Және де бұл Ережеде қамтылмаған немесе сол тілдің ерекшелігі болып табылатын грамматикалық категорияларды қосымша шартты белгілермен толықтыруға болады. Мұндай жағдайда көптеген тілдердің тәжірибесінде сол тілдің лексика-морфологиялық белгілерін халықаралық шартты белгілермен, оның мәнімен, сол тілдегі терминдік атауымен және мысалдармен көрсете отырып, олардың семантикалық белгілерін де қалыс қалдырмайды.

Осындай теггер жүйесін жасау мәселесі қазақ лингвистикасында қалай шешемін тауып жатыр деген сұраққа жауап беру үшін септіктердің теггік белгілеріне назар аударып көрдік. Қазақ тілінің Ұлттық корпусы мен публицистикалық корпустағы септік категориясының шартты белгілері мынадай: атау септік (ØAC/Nom), ілік септік (IC/Gen), барыс септік (БС/Dat), табыс септік (ТС/Acc), жатыс септік (ЖС/Loc), шығыс септік (ШС/Abl), көмектес септік (КС/Ins). Яғни дәстүрлі грамматикадағы септік атаулары мен тұлғаларынан ешқандай өзгешелік байқалмайды. Алайда мәтінді автоматты өңдеу үдерісі әрбір грамматикалық мағынаны есепке алу арқылы жүзеге асыралатындықтан, септіктің түрі бұрынғыдан көп болуы мүмкін. Мысалы, орыс тілінде мәтінді автоматты өңдеуге ұсыну үшін 70-ке жуық синтаксистік қатынас анықталып, сол бойынша зерттеу, талдау жұмыстары жүргізілуде. Оны түркі тілдерінің ерекшелігін ескере отырып зерттеулер жүргізуге арналған инструментарийлық бағдарлама болып табылатын «Тюркская морфема» порталында [18] қазақ тілінің дәстүрлі грамматикасында ескеріле бермейтін төмендегідей септік түрлері де граммемалар тізіміне қамтылғандығынан да аңғаруға болады:

№	Типологиялық атаулары		Белгіленім идентификаторы
	орысша	ағылшынша	
1	Совместный падеж	Commutative	COMIT
2	Сравнительный падеж	Comparative case	COM_CASE
3	Консекутив	Consecutive	CONSEC
4	Дательно-винительный падеж	Dative/Accusative	-DIR+ACC
5	Дательно-направительный падеж	Dative/Directive	DAT/DIR
6	Изъяснительный падеж	Deliberative	DELIB
7	Изъяснительный падеж+ притяжательный аттрибутивизатор	Deliberative+Possessive Attributivizer	DELIB+ATTR_POSS
8	Причинно-целевой падеж	Destinative	DEST
9	Направительный падеж	Directive	DIR
10	Инструментально-совместный падеж	Instrumentative/ Comitative	INST/COMIT
11	Ограничительный падеж	Limitive	LIM
12	Причинно-следственный падеж	Motivative	MOTIV
13	Ориентационный падеж	Orientative	ORIENT
14	Частный падеж	Partitive	PRTV
15	Продольный падеж	Prosecutive	PROS
16	Предельный падеж	Terminative	TERM

Кесте 3. «Тюркская морфема» порталындағы септік граммемалары

Нәтиже. Компьютер лингвистикасының, лингвистикалық типологияның заманауи зерттеулері тілдің құрылымдық-функционалдық ерекшеліктерін есепке алуды көздейді. Осыған орай тұлғалану формалары бір болса да, білдіретін мағынасы тіркесім табиғаты, мәнмәтіндік жағдайына қарай әр түрлі граммемалар пайда болуы мүмкін. Сондықтан әрбір граммеманы түгендеп, оған тиесілі шартты белгілерді бекіту қажет. Басқаша айтқанда, бір морфема арқылы екі түрлі граммема көрініс табатын жағдай болады [19]. Осы тұрғыдан келгенде, қазақ тілінің көмектес септігінің грамматикалық тұлғасын құралдық, біргелік, амалдық граммемаларға жіктеуге болады. Сондықтан көмектес септігін тек қана «КС/ Ins» шартты белгілерімен беру жеткіліксіз. Өйткені корпустарда омонимдерді ажырату, синтаксистік-семантикалық белгіленім жасау әрбір грамматикалық мағынаның нақты көрсетілуін қажет етеді.

Сонымен қатар қазіргі салғастырмалы-типологиялық және салыстырмалы-тарихи зерттеулерде тілдік материалдарды дәйек ретінде ұсынғанда бұрынғыдай ұзынсонар сөйлем ретінде емес, Лейпциг ережесіне негізделген жолма-жол морфемдік глостау («аудару») тәсілі қолданылады. Мұнда кез келген тілдің грамматикалық тұлғасы барынша дәлме-дәл беріліп, лингвистғалымдар басқа тілдің материалын «оқып», түсіне алатындай болуы көзделеді. Ал синтаксистік белгіленім – лексикалық бірліктер мен синтаксистік конструкциялар арасындағы байланысты сипаттайтын белгіленім түрі. Ол морфологиялық талдауыш мәліметтердің негізінде жүзеге асырылады. Яғни морфологиялық теггеудің шартты белгілері синтаксистік белгіленімге пайдаланылады деген сөз. Лингвистикалық формализм бағытында да сөйлемдерді формалдаудың метатілі мен сипаттау моделі кеңінен қолданылады. Бұл бағыт семантикаға көп көңіл бөледі және оны шартты семантикалық немесе категориалдық белгіленім деп те атайды. Ал «Сөз таптарының өзіне тән семантикалық, қасиеттері болғанымен,

тіркесіп келген сөздері немесе олардың аффикстерінің морфологиялық сипатына қарай мәні анықталып жататын жағдай да болады» [20, 149]. Осының барлығын есепке алу синтаксистік-семантикалық парсердің міндеті болып табылады. Егер де қазақ тілінің корпусын морфологиялық, синтаксистік-семантикалық және т.б. белгілері бойынша аннотациялаймыз десек, онда «... лингвистикалық параметрлер түрінде құрылымдалған, табиғи тілді автоматты өңдеуді (белгіленім жасауды) қамтамасыз ететін арнайы формальды белгілердің көмегімен көрініс табатын мәліметтер болуы керек» [21, 212]. Мұндай мәліметтерсіз табиғи тіл деректерін өңдеу жартылай немесе үстірт жасалған әрекет болып табылады.

Түйіндей келгенде, халықаралық стандартқа бағына отырып Лейпциг ережесі бойынша глостау, морфологиялық тегтеу, синтаксистік-семантикалық белгіленім жасау – мұның барлығы да морфологиялық категорияларға, соған басыбайлы етілген шартты белгілерге негізделеді. Сондықтан төл тіліміздің толыққанды Ұлттық корпусын әзірлейміз десек, ана тіліміздің бүге-шігесіне дейін «өрнектеуге» қауқары бар шартты белгілер жүйесін біріздендіріп, бекітіп алу – күн тәртібінде тұрған мәселенің бірі.

Қорытынды. Мемлекеттік тілдің және түркі тілдерінің корпустарында немесе түркі тілдерінің метатілін жасауды мақсат тұтқан отандық компьютер лингвистикасы саласындағы мамандар ұсынып жүрген морфологиялық шартты белгілерде бірізділікжоқнемесеәлідетолықтыруларды қажет етеді. Қазіргі қолданыстағы шартты белгілер (тегтер) тіліміздің синтаксистік-семантикалық картинасын бере алатындай дәрежеге жеткен жоқ.

Заманауи лингвист-ғалымдар кез келген тілдің типологиялық ерекшелігін немесе барлық тілдерге тән әмбебап құбылыстарды ажыратып тану үшін, зерттеулерде лингвистикалық сипаттама беру үшін глостаудың халықаралық стандартын білуі керек. Және де аталған стандарт негізінде ана тіліміздің табиғатын толық таныта алатын шартты белгілер жүйесін жасақтап алуымыз қажет.

Дәстүрлі лингвистикалық зерттеулердегі грамматикалық категориялардың топтастырылуын қайтадан қарау қажеттілігі туындап отыр. Себебі табиғи тілдің деректерін машина түсінетіндей болуы үшін формалдаудың құрылымдық және семантикалық мән-мағынасы түрлендіріліп ұсынылуы тиіс. Мысалы, атау септігі мен табыс септігінің нөлдік тұлғасын автоматты түрде ажырату үшін олардың шартты белгілері бірдей болмасы түсінікті.

Қазақ тілінің, одан әрі түркі тілдерінің шартты белгілері жүйесін әзірлеуде бірізділікті сақтау үшін компьютер лингвистикасы саласы мен тіл мамандары бірлесе отырып шешім қабылдағаны жөн. Жоғарыда сөз еткеніміздей, олардың табиғи тілді пайдаланудағы мақсаттары әр түрлі болғанымен, екі саланың мамандары да тілдік деректерді машинаға автоматты түрде өңдеуге ұсынатындықтан, олардың мүдделерін түйістіретін тұсы осы болмақ.

Әдебиеттер тізімі

1. Абдикарим Н. Қазақ тілінің морфосемантикалық және синтаксистік метабелгілер жүйесін әзірлеу мәселесіне // Қазақ тілі грамматикасының қазіргі зерттеу парадигмалары және оқытудың инновациялық технологиялары: филология ғылымдарының докторы, профессор Т. Сайрамбаевтың 85 жылдығына арналған халық. ғыл.-әдіст. конф. материалдары. – Алматы, 2022. – Б. 83-85.

2. Орехов Б. В., Резникова Т. И. Компьютерные перспективы лексико-типологических исследований // Вестник ВГУ. Серия: Лингвистика и межкультурная коммуникация. – 2015. – № 3. – С. 17-23.

3. Қазақстан Республикасындағы тіл саясатын іске асырудың 2020 - 2025 жылдарға арналған мемлекеттік бағдарламасын бекіту туралы Қазақстан Республикасы Үкіметінің 2019 жылғы 31 желтоқсандағы № 1045 қаулысы [Электрондық ресурс]. – 2019. – URL: https://adilet.zan.kz/kaz/docs/P1900001045 (қаралған күні: 19.05.2022).

4. Пірманова К.Қ., Жаңабекова А.Ә., Барменқұлова А. Ұлттық корпустарға негізделген лингвистикалық зерттеулер жүргізу (қазақ, орыс, ағылшын тілі материалдары негізінде) // Әл-Фараби атындағы қазақ ұлттық университеті. Филол. сериясы. – 2022. – №3. - Б. 83-93.

5. Светлов А.В., Комендантов А.С. Автоматизация процесса получения лингвистической информации: современные возможности // Вестник ВолГУ. Сер. языкозн. – 2017. – Т. 16. № 2. – С. 39-46.

6. Шарипбай А.А., Гатиатуллин А.Р., Ергеш Б.Ж., Қажымұхан Д.А. Разработка единого метаязыка морфологии тюркских языков // Вестн. КазНУ. Сер. мат., мех., инф. – 2018. – № 4. – С. 78-87.

7. Қазақ тілінің ұлттық корпусы [Электрондық ресурс]. – URL: https://qazcorpus.kz/ (қаралған күні: 03.12.2022).

8. Қазақ тілі ұлттық корпусының кіші корпустары [Электрондық ресурс]. – URL: https://qazcorpora.kz/ (қаралған күні: 03.12.2022).

9. Torotoev G., San. G. Torotoeva. Linguistic Annotation of Grammatical Categories of Sakha: Nouns Gavril. Journal of Siberian Federal University. Humanities & Social Sciences. – 2018. – P.1-10.

10. Национальный корпус русского языка. [Электронный ресурс]. – URL: https://ruscorpora.ru/page/instruction-morph/ (дата обращения: 21.10.22).

11. Касевич В. Б. Введение в языкознание: учеб. для студ. учреждений высш. проф. образования / В. Б. Касевич. — 2-е изд., испр. и доп. — СПб.: Филологический факультет СПбГУ; М.: Издательский центр «Академия», 2011. – 240 с.

12. Галичкина Е.Н. Компьютерная коммуникация: лингвистический статус, знаковые средства, жанровое пространство: автореф. дис. д-ра филол. наук: 10.02.19 / Е.Н. Галичкина; Волгоград. гос. соц.-унив. – Волгоград, 2012. – 40 с.

13 Электронный корпус татарского языка [Электронный ресурс]. – URL: http://tatmorphan.pythonanywhere.com/morphan_tags (дата обращения: 12.12.2018).

14. Орехов Б.В. Проблемы морфологической разметки башкирских текстов // Труды Казанской школы по компьютерной и когнитивной лингвистике. – Казань: «Фэн» Академии наук РТ. – 2014. – С. 135-140.

15. List of glossing abbreviations. Available at: https://en.wikipedia.org/wiki/List_of_glossing_abbreviations. (accessed: 05.12.2022).

16. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. Available at: https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf (accessed: 12.10.2021).

17. Восточно-армянский национальный корпус [Электронный ресурс]. – URL:. http://eanc.net/EANC/search/?interface_language=ru (дата обращения: 07.12.2022).

18. Портал «Тюркская морфема»: граммемы [Электронный ресурс]. – URL: http://modmorph.turklang.net/ru/grammeme (дата обращения: 05.12.2022).

19. Словарь иностранных слов русского языка [Электронный ресурс]. – URL: https://dic.academic.ru/dic.nsf/dic_fwords (дата обращения: 12.12.2018).

20. Daniel J., James H. Martin. Speech and Language Processing. – 2020. – P. 623. Available at: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf (accessed: 18.09.2022).

21. Сулейманов Д.Ш., Хакимов Б.Э., Гильмуллин Р.А. Корпус татарского языка: концептуальные и лингвистические аспекты // Вестник ТГПГУ. – 2011. - №4 (26). – С. 211-216.

References

1. Abdikarim N. Qazaq tilinin morfosemantikalyq zhane sintaksistik metabelgiler zhujesin azirleu maselesine [Morphosemantic of the Kazakh language and to the problem of developing a system of syntactic metasigns]. T.Sajrambaevtyn 85 zhyldygyna arnalgan halyq. gylymj-adist. konf. materialdary [Materials of International scientific and methodological conference «Modern research paradigms of Kazakh grammar and innovative teaching technologies» dedicated to the 85th birthday anniversary of the doctor of philology, professor T.Sayrambaev] (Almaty, 2022, pp. 83-85). [in Kazakh]

2. Orekhov B. V., Reznikova T. I. Komp'yuternye perspektivy leksiko-tipologicheskih issledovanij [Computer prospects for lexico-typological research], Vestnik VGU. Seriya: Lingvistika i mezhkulturnaya kommunikatsiya [Bulletin of VSU. Series: Linguistics and Intercultural Communication], 3, 17-23 (2015). [in Russian]

3. Qazaqstan Respublikasyndagy til sayasatyn iske asyrudyn 2020 - 2025 zhyldarga arnalgan memlekettik bagdarlamasyn bekitu turaly Qazaqstan Respublikasy Ykimetinin 2019 zhylgy 31 zheltoksandagy № 1045 qaulysy [Resolution of the Government of the Republic of Kazakhstan "On approval of the state program for the implementation of language policy in the Republic of Kazakhstan for 2020-2025" dated December 31, 2019, No. 1045]. Available at: https://adilet.zan.kz/kaz/docs/P1900001045 (accessed: 19.05.2022).

4. Pirmanova K.K., Zhanabekova A.A., Barmenkulova A. Ulttyq korpustarga negizdelgen lingvistikalyq zertteuler zhurgizu (qazaq, orys, agylshyn tili materialdary negizinde)) [Conducting linguistic research based on national corpora (based on Kazakh, Russian, English materials)], Al-Farabi atyndagy kazak ulttyk universiteti. Filol. Serijasy [Bulletin of Al-Farabi Kazakh National University. Philol. Series], 3, 83-93 (2022). [in Kazakh]

5. Svetlov A.V., Komendantov A.S. Avtomatizatsiia protsessa polucheniia lingvisticheskoi informatsii: sovremennye vozmozhnosti [Automation of the process of obtaining linguistic information: modern possibilities], Vestn. VolGU. Ser. iazykozn. [Bulletin of VolSU. Ser. Linguistic], 16 (2), 39-46 (2017). [in Russian]

6. Sharipbai A.A., Gatiatullin A.R., Ergesh B.J., Kazhymukhan D.A. Razrabotka edinogo metaiazyka morfologii türkskih iazykov [Development of a unified metalanguage of the morphology of the Turkic languages], Vestn. KazNU. Ser.: mat., mech., inf. [Bulletin of Al-Farabi KazNU. Ser. mat., meh., inf.], 4, 78-87 (2018). [in Russian]

7. Qazaq tilinin ulttyq korpusy [National corpus of the Kazakh language]. Available at: https:// qazcorpus.kz/ [in Kazakh] (accessed 03.12.2022).

8. Qazaq tilinin ulttyq korpusynyn kishi korpustary [The small cases of the national corpus of Kazakh language]. Available at: https://qazcorpora.kz (accessed: 03.12.2022). [in Kazakh].

9. Torotoev G., San. G. Torotoeva. Linguistic Annotation of Grammatical Categories of Sakha: Nouns Gavril. Journal of Siberian Federal University. Humanities & Social Sciences. – 2018. – P.1-10.

10. Natsionalnyi korpus russkogo yazyka [Russian National Corpus], Available at: https://ruscorpora. ru/page/instruction-morph/ (accessed: 21.10.22). [in Russian]

11. Kasevich V. B. Vvedenie v yazykoznanie [Introduction to linguistics] (Akademiya, Saint Petersburg, 2011).

12. Galichkina E.N. Kompyuternaya kommunikatsiya: lingvisticheskii status, znakovye sredstva, zhanrovoe prostranstvo [Computer Communication: Linguistic Status, Symbolic Means, Genre Space]. Autoref. dis. Doctor of Philology: 10.02.19 (Volgograd, 2012. 40 p.).

13. Elektronnyi korpus tatarskogo yazyka [Electronic corpus of the Tatar language]. Available at: http://tatmorphan.pythonanywhere.com/morphan_tags (accessed: 12.12.2018). [in Russian]

14. Orekhov B.V. Problemy morfologicheskoi razmetki bashkirskikh tekstov [Problems of morphological marking of Bashkir texts], Trudy Kazanskoi shkoly po kompyuternoi i kognitivnoi lingvistike [Proceedings of the Kazan school on computer and cognitive linguistics], 135-140 (2014). [in Russian]

15. List of glossing abbreviations. Available at: https://en.wikipedia.org/wiki/List_of_glossing_ abbreviations. (accessed 05.12.2022).

16. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. Available at: https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf (accessed: 12.10.2021).

17. Vostochno-armyanskii natsionalnyi korpus [Eastern Armenian National Corpus], Available at: http://eanc.net/EANC/search/?interface_language=ru (accessed: 07.12.2022). [in Russian]

18. Portal «Tyurkskaya morfema»: grammemy [Portal “Turkic morpheme”: grammes]. Available at: http://modmorph.turklang.net/ru/grammeme (accessed: 05.12.2022). [in Russian]

19. Slovar inostrannykh slov russkogo yazyka [Dictionary of foreign words of the Russian language]. Available at: https://dic.academic.ru/dic.nsf/dic_fwords (accessed: 12.12.2018). [in Russian]

20. Daniel J., James H. Martin. Speech and Language Processing. – 2020. – P. 623. Available at: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf (accessed 18.09.2022).

21. Suleimanov D.Sh., Khakimov B.E., Gilmullin R.A. Korpus tatarskogo yazyka: kontseptualnye i lingvisticheskie aspekty [Corpus of the Tatar language: conceptual and linguistic aspects], Vestnik TGPGU [Bulletin of the TSPSU], 4(26), 211-216 (2011). [in Russian]

Н. Абдикарим

Национальный научно-практический центр «Тіл-Қазына» имени Ш.Шаяхметова, Астана, Казахстан

Предпосылки создания синтаксической и семантической разметки

Аннотация. В статье рассматривается дальнейшее совершенствование корпуса казахского языка, которое развивается в связи с политикой цифровизации страны, его использования в лингвистических исследованиях, учебном процессе и т.д. Необходимо создать и официально утвердить перечень условных признаков морфологического глоссирования на основе международного Лейпцигского правила. Проведен анализ использования метатегов, используемых в корпусах казахского языка и в сфере отечественной компьютерной лингвистики. Также освещены работы родственных и других языков в этом направлении, показаны их специфика и основные характеристики условных знаков и обозначений. Условные знаки, используемые при морфологическом глоссировании, могут стать основой для создания корпуса казахского языка, оснащенного синтаксически-семантическими разметками. Такие условные знаки дают возможности нашему родному языку сравнивать или сопоставлять его с другими языками, продемонстрировать языковые материалы в кратком и наглядном варианте, описать грамматические признаки, обрабатывать текстовые данные на компьютере; они открывают новые возможности для современных лингвистических исследований с использованием IT-технологий и разработанных корпусов.

Ключевые слова: политика оцифровки, языковой корпус, лейпцигское правило, глоссирование, морфологический тег, условные знаки, синтаксически-семантическая разметка, современные лингвистические исследования.

N. Abdikarim

National Scientific and Practical Center named after Sh. Shayakhmetov, Astana, Kazakhstan

Prerequisites of syntactic-semantic markings

Abstract. The article discusses the issue of the need to create and officially approve a list of conventional signs (marks) of morphological glossing on the basis of the international Leipzig rule for further improvement of the Kazakh language corpora, developed in connection with the country’s digitalization policy, and its use in linguistic research, educational process and other purposes. The analysis of meta tags used in the corpora of the Kazakh language and in the field of domestic computational linguistics has been carried out. At the same time, attention is also paid to the work of related languages in this direction and it is shown that the conventional signs and symbols in them have common features or features that are unique to this language. According to the author, the conventional signs used in morphological glossing are the basis for creating a corpus of the Kazakh language, equipped with syntactic-semantic markup. Also, such conventional signs provide new opportunities for comparative or comparative study of the native language with other languages as a reference language, for the presentation of language materials in a concise and visual form, for the description and processing of text data on a computer; and new opportunities for conducting modern linguistic research using IT technologies and developed corpora.

Keywords. Digitization policy, language corpus, Leipzig rule, glossing, morphological tag, conventional signs, syntactic-semantic notation, modern linguistic research.

Автор туралы мәлімет:

Абдикарим Н. – филология ғылымдарының кандидаты, Терминология басқармасының жетекші ғылыми қызметкері, Ш. Шаяхметов атындағы ұлттық ғылыми-практикалық орталығы, Астана, Қазақстан.

Abdikarim N. – Candidate of Philological Sciences, Leading Researcher, Department of Terminology, National Scientific and Practical Center named after Sh. Shayakhmetov, Astana, Kazakhstan.