ҒТАМР 16.31.21

А.Ә. Жаңабекова, А.Қ. Қожахметова*

А.Байтұрсынұлы атындағы Тіл білімі институты, Алматы қ., Қазақстан
*E-mail: akony_8484@mail.ru

МЕТАБЕЛГІЛЕНІМ ЕНГІЗІЛГЕН МӘТІНДЕРДІ АРНАЙЫ КОМПЬЮТЕРЛІК БАҒДАРЛАМА БОЙЫНША ӨҢДЕУ

Мақалада қазақ тілінің ұлттық корпусының құрамындағы мәтіндерді металингвистикалық белгілеу мәселесі қарастырылған. Ұсынылған қазақ мәтіндерінің метабелгілеу жүйелері сипатталған. Метабелгілеуде мәтінді тұтасымен сипаттайтын параметрлер мәтініне сілтеме жасау жатады. Мета-таңбалау, қолданушының көрсетілген қасиеттері бар ішкі құрылымды құрастыру үшін мәтіндерді іздеу және таңдау мүмкіндігін береді. Метабелгіленімдер – ғылыми-зерттеу жұмыстарында белгілі бір кезеңге, стильге, авторға, тақырыпқа т.б. қатысты материалдар жинаудың таптырмас дереккөзі десек те болады. Тіл корпусын жасауда әдебиеттер таңдауда мәтіндердің жанрлық түрлілігін ескеру ең басты міндеттердің бірі саналады. Корпус жасауда метабелгіленімдерді енгізу қажеттігі дау тудырмайды. Олар корпусты пайдаланушыға, әсіресе тілші-зерттеушілерге кез келген қажетті ақпаратын тез әрі оңай тауып алуға мүмкіндік береді.

Сонымен қатар лингвистикалық ақпараттар мүмкіндігінше барлық тіл деңгейлерінен алынады. Бірақ корпус жасауда бірден барлық деңгей бойынша лингвистикалық ақпарат беру аса күрделі әрі мүмкіндігі де шектеулі. Солай бола тұра, «Қазақ тілінің мәтіндер корпусында» метамәтіндік белгіленімдер мен лингвистикалық белгіленімдерден – морфологиялық, сөзжасамдық, фонетикалық, морфосемантикалық белгіленімдер программасы іске қосылған. Мақалада корпусқа метабелгіленімдердің түрлері (23 параметр), лингвистикалық белгіленімдер программасының жұмыс істеу механизмдері туралы сипатталады.

Тірек сөздер: корпустық лингвистика, мәтіндер корпусы, белгіленім, метабелгіленім.

А.А. Жанабекова, А.К. Кожахметова

Институт языкознания имени А.Байтурсынулы, Алматы, Казахстан
*E-mail: akony_8484@mail.ru

ОБРАБОТКА ТЕКСТОВ, ВКЛЮЧЕННЫХ В МЕТАРАЗМЕТКУ, НА СПЕЦИАЛЬНОМ КОМПЬЮТЕРНОМ ПРОГРАММНОМ ОБЕСПЕЧЕНИИ

В статье рассматривается проблема металингвистической разметки текстов в составе Национального корпуса казахского языка. Описаны метаразметные системы казахских текстов. Метаразметка включает в себя ссылку на текст поисковым параметрам, характеризующих фрагмент в целом. Мета-разметка позволяет пользователю искать

и выбирать тексты для составления внутренней структуры с указанными свойствами. Метаразмеченные материалы – незаменимый источник в научно-исследовательской работе по сбору текстового контента, относящегося к определенному периоду, стилю, автору, теме и др. Одной из главных задач при выборе литературы при создании корпуса является учет жанрового разнообразия текстов. Необходимость введения метаразметок при создании корпуса не вызывает споров. Они позволяют пользователю корпуса, особенно исследователям-корреспондентам, быстро и легко находить любую необходимую информацию.

Кроме того, лингвистическая информация по возможности извлекается со всех языковых уровней. А вот в корпусах предоставление лингвистической информации сразу по всем уровням крайне сложно и возможности ограничены. Тем не менее, в «корпусе текстов казахского языка» запущена программа метатекстовых разметок и лингвистических разметок – морфологических, словообразовательных, фонетических, морфосемантических разметок. В статье описываются виды метатекстов на корпус (23 параметра), механизмы функционирования программы лингвистических разметок.

Ключевые слова: Корпусная лингвистика, корпус текста, разметка, метаразметка.

A.A. Zhanabekova, A.K. Kozhahmetova

A. Baitursynuly Institute of Linguistics, Almaty, Kazakhstan.
*E-mail: akony_8484@mail.ru

PROCESSING OF THE TEXTS INCLUDED IN METAMARKING ON THE SPECIAL COMPUTER SOFTWARE

The article deals with the problem of metalinguistic markup of texts within the National Corpus. The meta-markup systems of the proposed Kazakh texts are described. Meta-markup includes a link to the text of parameters that characterize the text as a whole. Meta-markup, allows the user to search and select texts to compose an internal structure with specified properties. Meta-tagged materials are an indispensable source in research work on the collection of materials related to a certain period, style, author, topic, etc. One of the main tasks when choosing a literature when creating a corpus is taking into account the genre diversity of texts. The necessity of introducing meta-markings in the manufacture of the case is not controversial. They allow the user of the corpus, especially the correspondent researchers, to quickly and easily find any information they need.

In addition, linguistic information is extracted whenever possible from all language levels. But in the corpus, the provision of linguistic information at all levels at once is extremely difficult and the possibilities are limited. Nevertheless, a program of metatext markings and linguistic markings – morphological, word-formation, phonetic, morpho-semantic markings, has been launched in the "corpus of Kazakh texts". The article describes the types of metatext on the corpus (23 parameters), the mechanisms of functioning of the program of linguistic markup.

Key words: Corpus linguistics, text corpus, annotation, meta-markup.

Кіріспе

XX ғасырда басталған ғылыми-техникалық «революция» әлемнің кез келген

мемлекетінің ішкі-сыртқы саясатына, әсіресе экономикалық әлеуетіне ерекше серпін беріп қана қоймай, Тәуелсіз Қазақстан Республикасы сияқты дамушы елдердің жас мемлекет ретінде қалыптасуында айрықша рөл атқарды. Қоғамдық қызметтің қай саласында да қолданбалы бағыт басымдық алды. Осы орайда лингвистиканың қолданбалы саласы да қалыптасып, дәстүрлі тіл білімінің бағыттарын өз әдіс-тәсілдерімен зерттеуге кірісті.

Қазақ тіл білімінің қолданбалы лингвистика саласы бойынша 70-жылдары тілдік материалдарды автоматты өңдеу мәселесіне байланысты математикалық, формалды-логикалық, статистикалық-ықтималдық әдістерді игеру қажеттігі туды. Соның нәтижесінде қазақ тіл білімінде математикалық лингвистика саласы дүниеге келіп, тілдік бірліктерді модельдеуге қатысты докторлық, кандидаттық диссертациялар қорғалды және жас ғалымдардың монографиялары жарияланды.

Бұрын негізінен статистика мен лексикографиялық зерттеулер бойынша қолданылған тілді автоматтандыру мәселесі тіл білімінің жекелеген салалары үшін де қажеттілік тудырды. Бүкіл тіл жүйесін біртіндеп компьютер жадына енгізіп, оларды белгілі бір автоматты басқару орталығы арқылы іске қосу, әсіресе, тіл тұтынушыларының практикалық қажеттілігі үшін өте маңызды. Басқаша айтқанда, ғылым-білім саласындағы қоғамдық-әлеуметтік сұраныстарды жеңіл әрі тез өтейтін бірден-бір ақпараттық жүйе ретіндегі орны ерекше.

Қазіргі кезде компьютерлік лингвистика саласы бойынша осындай ақпараттық сұраныс жүйесін қалыптастыру мақсатында «тілдік корпустарды» зерттеу мен оларды қазақ тілі материалдары негізінде қолданысқа енгізу мәселесі қызығушылық тудыруда.

Компьютерде мәтіндерге лингвистикалық ақпараттар беру, корпустық зерттеулер біраз уақыттан бері жалпы тіл білімінде, соның ішінде шетел тіл білімінде лингвистикалық зерттеулердің негізгі тәсілдерінің бірі ретінде қолға алынып келеді. Мұндай зерттеу жұмыстары қазақ тіл білімінде бұрын-соңды зерттеу нысанына ілінген жоқ, тіпті орыс тіл білімінің өзінде бұл мәселе лингвистиканың басқа салалары мен бағыттарына қарағанда кейіндеп қалған.

Қазіргі жаһандану кезеңінде әртүрлі саяси-әлеуметтік, экономикалық қарым-қатынастарға байланысты ақпарат ағыны бұрын-соңды болмаған қарқынмен өршуде. Ал қоғам өміріндегі мұндай ақпарат ағымының таралуы табиғи тілде жүзеге асатындықтан, тіл білімінің қызметі күннен-күнге кеңеюде. Осыған байланысты ұшы-қиырсыз ақпарат ағынын игеру мақсатында шетел және орыс тіл білімінде орасан зор нәтиже беріп отырған тілдік корпустарды қазақ тіл білімінің материалдары негізінде жасау бүгінде үлкен сұранысқа ие болып отыр. Осы ретте қазақ әдеби тіліндегі әртүрлі функционалдық стильдерді (көркем проза, драматургия, ғылыми-техникалық мәтіндер, публицистикалық мәтіндер) қамтитын лингвистикалық ақпаратпен толық аннотацияланған (мазмұндалған) ұлттық корпусты құрастыру – бүгінгі таңда аса өзекті мәселе.

Тілдік материалдарды автоматтандыру қажеттігі қазіргі кезде тіл білімінің «Компьютерлік лингвистика» атты саласының дамып, қалыптасуына мүмкіндік берді. Тіл білімінің жаңа саласы болып табылатын «Корпустық лингвистика» – «Компьютерлік лингвистиканың» бір саласы.

Корпустық лингвистика – қолданбалы лингвистиканың заман сұранысынан туындаған салаларының бірі. Корпустық лингвистика тілді ақпараттандыру, техника

мен технологияны пайдалана отырып зерттеуді қажет ететін маңызды ғылым саласы болып табылады.

Материалдар мен әдістер

Корпустық лингвистика компьютерлік технологияларды қолдана отырып, лингвистикалық корпустар (мәтіндер корпустарын) құрастыру мен оны пайдаланудың жалпы ұстанымдарын зерттейді. Ал лингвистикалық немесе тілдік, мәтіндер корпусы дегеніміз – нақты тілдік мәселелердің шешімін табуға арналған аса үлкен көлемдегі машине (компьютер) оқи алатындай түрде көрініс табатын, бірыңғайланған, құрылымдалған, белгіленген (шартты белгілер қойылған), филологиялық тұрғыда компетентті саналатын тілдік деректер ауқымы. Басқаша айтқанда, қазақ тілінің мәтіндер корпусы (тілдік корпусы) – аса ірі көлемдегі жүйелі түрде құрылымдалған, тиісті белгіленімдер енгізілген нақты тілдік мәселелерді компьютер арқылы шешуге арналған тілдік деректер ауқымы (массиві).

Қазіргі кезде «корпус» ұғымының бірнеше анықтамалары бар. Мысалы, Э.Финеганның оқулығында мынадай анықтама берілген: «корпус – мәшине (компьютер) оқи алатын форматта және мәтін туралы ақпарат берілген, яғни айтушы, автор, адресат (тыңдаушы) немесе аудитория (тыңдаушылар) жайлы ақпараттар қамтылатын мәтіндердің репрезентатты (көлемді) жиыны» (Finegan, 2004).

Википедия желісі корпустарға статистикалық талдау мен болжамдардың шынайылығын, белгілі бір лингвистикалық салалар бойынша тілдік ережелердің негізділігін тексеру үшін қолданылатын үлкен көлемді және құрылымдалған электронды пішіндегі мәтіндер жиынтығы деген анықтама береді (Wikipedia). Т. Мак Энери және Э. Вилсон мынадай анықтама ұсынады: корпус – ол тілдік модель ретінде қолданылатын нақты тілдік критерийлерге сәйкес келетіндей таңдалып алынған тілдік фрагменттердің жиынтығы (McEnery T., Wilson, 2001). Ғалым В.В. Рыков мәтіндер корпусын, негізінде логикалық түпкі ой, логикалық идея жатқан мәтіндер жиынтығы деп түсіндіреді де, ал олар мәтіндерді құрастыру ережелері, мәтіндерді талдаудың алгоритмі мен бағдарламасы және осымен түйіндес идеология мен әдіснамалар арқылы жүзеге асады, – дейді (Рыков, 2002).

В.П.Захаровтың корпус ұғымын тар мағынада түсіндіруі қазіргі ғылыми түсінікте эволюциялық тұрғыда көрініс тапқан. Мысалы, ғалым «Корпусная лингвистика» атты оқу құралында былай дейді: «Под названием лингвистический, или языковой, корпус текстов понимается большой, представленный в электронном виде, унифицированный, структурированный, размечанный, филологический компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» (Захаров, 2005).

Сонымен, корпус дегеніміз – әр тілдегі электронды пішінге келтірілген, яғни бір басқару орталығынан автоматты түрде жұмыс істейтін, лингвистикалық ақпараттарды қамтитын мәтіндер жинағы.

Мәтіндер корпусын сөз етер алдында, ең алдымен, мәліметтер корпусы ұғымын түсініп алу қажет болады. Мәліметтер корпусы дегеніміз – феномендері лингвистикалық сипаттауды қажет ететін, белгілі ережелерге сай құрастырылған және тілдік жүйеде іске асатын деректер таңдамасы. Ол тек бірөлшемді – тілдік (речевое).

Корпустар атқаратын функциясына қатысты түрлі шағын топтарға (подкорпус) бөлінеді. Әсіресе лингвистикалық зерттеулер үшін лингвистикалық белгіленімдер

қойылған аннотацияланған корпустар құрастыру өте маңызды әрі пайдалы. Осындай белгіленім қойылу-қойылмауына қарай корпустар белгіленім қойылған (аннотацияланған) және белгіленім қойылмаған (аннотацияланбаған) болып екіге бөлінеді. Белгіленім қойылған корпустар құрастыру лингвистикалық зерттеулер үшін маңызды болғанмен, ең алдымен корпусқа енгізілетін мәтіндерді сұрыптаудан өткізіп, олар туралы нақты ақпараттар беру қажет. Мұны кейде экстралингвистикалық белгіленімдер деп те атайды, метамәтіндік белгіленім (метатекстовая разметка – метаразметка) деп те қолданылады. Лингвистикалық белгіленімдер ішкі белгіленім деп аталса, метамәтіндік белгіленімдерді сыртқы белгіленім деп бөлетіндер де бар. Олар: мәтін және автор туралы ақпарат: автор, атауы, жылы, шыққан жері, мәтіннің жанры, тақырыбы, стилі, көлемі т.б. Оларды библиографиялық, типологиялық, тақырыптық, әлеуметтік, формальдық (мәтін, тарау, бөлім, абзац, сөйлем т.б.) және техникалық (орындаушылар, электрондық нұсқа алынған дереккөз, өңделген күні, кодталған кезі т.б.) деп те бөледі.

Экстралингвистикалық белгіленім немесе метадеректер қатарына «сыртқы», «интеллектуалдық» белгіленімдер (библиографиялық сипаттама, типологиялық сипаттама, тақырыптық сипаттамалар, әлеуметтік сипаттамалар), «формалды» құрылымдық белгіленімдер (мәтін, бөлім, тарау, тарауша, абзац, сөйлем), сонымен бірге техника-технологиялық белгіленімдер (кодтау, өңдеу даталары, орындаушыларды, электрондық нұсқаның дереккөзі) жатады. Метадеректер жиыны мәтіндер корпустарындағы зерттеушілерге ұсынылатын мәліметтер мүмкіндіктерін айқындайды. Сондықтан деректерді таңдау кезінде зерттеудің мақсатын және тілшілердің сұранысын және мәтінге қосымша белгілерді енгізу мүмкіндіктерін басшылыққа алу қажет болады.

Метамәтіндік белгіленімдер корпустың қай түріне болмасын енгізілуге тиісті маңызды ақпараттар болып табылады, сонымен қатар лингвистикалық белгіленімдер қойылатын корпустар құрастыру жұмысынан бұрын, корпусқа әртүрлі стильдерден алынған мәтіндерді енгізу барысында атқарылады.

Зерттеу жұмыстарының белгілі бір жүйе бойынша жүргізілетіні белгілі. Өйткені тіл – қарым-қатынас құралы, қолданыс аясы өте кең күрделі жүйе болғандықтан, тілдік зерттеулер белгілі автор шығармалары бойынша немесе белгілі бір кезеңде жарық көрген көркем не баспасөз немесе тарихи ескерткіштер бойынша, сол сияқты белгілі бір жанр, стиль немесе белгілі бір тақырып, мәселе бойынша тарам-тарам болып жіктеліп кете береді. Осындай әртарапта шашылып жатқан материалдарды корпус жадына жай ғана енгізбей, белгілі бір жүйеге бағындырып енгізсе, зерттеу жұмыстарының да тиімді орындалуына ықпал етеді. Ал бұлайша корпус жадына енгізілген мәтіндерді жүйелі құрылыммен беру метабелгіленімдер қоюдың теориялық және практикалық әдіс-тәсілдерін меңгергенде ғана мүмкін болады. Сондықтан корпустың қайсібір түрінде болмасын, метабелгіленімдер қою мәселесі дұрыс жүргізілуі керек. Бұл да теориялық бағытта зерттеліп, әрі қолданбалы бағытта жүзеге асырылатын көлемді де, күрделі әрі бүгінгі жаһандану заманында кезек күттірмейтін өзекті тақырыптардың бірі. Қазақ тілінің ұлттық корпусы тиімді құрастырылып, өз дәрежесінде қолданысқа енгізілуі үшін әлемдік тәжірибедегі метабелгіленімдер қоюдың ортақ ұстанымдары мен әдіс-тәсілдерін зерттеу қажет.

Метабелгіленімдер – ғылыми-зерттеу жұмыстарында белгілі бір кезеңге, стильге, авторға, тақырыпқа т.б. қатысты материалдар жинаудың таптырмас дереккөзі. Мета-

белгіленімдер қойылған корпустардан зерттеушілер өзіне қажетті стиль, кезең, автор т.б. ақпараттар бойынша мәліметтерді тез тауып алуына мүмкіндік алады. Ал бұл жетістік қазіргі дамыған еліміздің ғылыми-зерттеушілік әлеуетін, ғылым мен білімді жан-жақты дамытатын бірден-бір күш екендігі сөзсіз.

Корпустың қайсібір түрінде болмасын, ең алдымен жинақталған мәтін туралы ақпарат беретін метамәтіндік параметрлер компьютерлік базаға салынуы қажет. Сондықтан корпус құрастыруда метабелгіленімдер (шығарма атауы, авторы, жылы, шыққан орны, стилі, жанры, көлемі т.б.) жасаудың әдіс-тәсілдерін, түрлерін, компьютерлік бағдарламаларын т.б. зерттеудің маңызы зор.

Әлемдік тіл білімінде метабелгіленім қоюдың теориялық ұстанымдары мен практикалық әдіс-тәсілдері бар. Орыс тілінің ұлттық корпусындағы метабелгіленімдер қоюдың теориялық, практикалық мәселелерін С.О.Савчук қарастырған (Савчук, 2005).

Корпусқа метабелгіленімдер енгізу жұмысынан бұрын атқарылатын жұмыс – әртүрлі стильге тән мәтіндердің электронды нұсқасын жинақтау болып табылады. Мәтіндер жинақтау жұмысын әртүрлі сипатта жүзеге асыруға болады. Олар: 1) интернет сайттарына салынған мәтіндерді қолдап іздеу арқылы WORD бағдарламасына көшіріп алу немесе PDF бағдарламасындағы мәтіндерді арнайы бағдарламалар арқылы WORD бағдарламасына көшіру; 2) интернет сайттарындағы мәтіндерді автоматты түрде пайдалана алатын бағдарламалар арқылы мәтін жинау; 3) кітап, газет мәтіндерін сканерлеу арқылы мәтін жинақтау; 4) баспалармен келісіп, әртүрлі мәтіндердің (көркем проза, поэзия, драматургия, газет-журнал, ғылыми-техникалық мәтіндер т.б.) электронды нұсқаларын алу; 5) ауызекі сөйлеу корпустары үшін таспаға жазылып алынған видео, аудио таспаларды телевидения, радио т.б. мекемелерден алу немесе ауызша тілдесімдерді таспаға арнайы жазып алу арқылы ауызша мәтіндер жинау т.б.

Корпуста алынатын мәтіндердің авторы, сол сияқты мазмұны, сапасы да ескеріледі. Қысқасы, корпуста белгілі бір кезеңдегі тіл қолданысын толықтай сипаттайтындай функционалды стильдер қамтылуы қажет.

Зерттеу нәтижелері мен талқылау

Корпусқа белгілі бір әдеби тілдегі әртүрлі стильдер, әртүрлі авторлар шығармалары мен еңбектері енгізілетіндіктен, тілтұтынушысының оларды белгілі бір жүйемен пайдалануы үшін олар туралы егжей-тегжейлі ақпарат беріледі. Корпустың іздеу жүйесіне салынатын мәтіндер туралы ақпараттар (метабелгіленімдер) әр тілде әртүрлі болып келеді. Мысалы, Орыс тілінің ұлттық корпусында (ОТҰК) мәтін туралы ақпараттар 25 параметрге негізделген. Олардың ішінде 9-ы мәтінді сипаттайтын мәліметтер, 3-еуі автор туралы, 3-еуі мәтіндердің бағытталған аудиториясы, 4-еуі мәтіндер туралы библиографиялық мәліметтер, 5-еуі қосымша ақпараттар.

Нақтырақ айтсақ:

Бірінші блок:

1. *мәтін авторы:* аты-жөні, жынысы, туған уақыты (немесе шамамен есептегендегі жасы);

2. *мәтіннің атауы;*

3. *мәтіннің құрастырылу уақыты мен орны* (дәл немесе шамамен көрсетуге болады);

4. *мәтін көлемі:* көркем шығармалар үшін мынадай көлемдер қабылданған: әңгіме 5 мың сөзден көп емес; әдеттегі повестер ұзындығы – 5 мың мен 15 мың сөз аралығы; романның әдеттегі ұзындығы – 15 мың сөзден асады.

Екінші блок: мәтіндер корпусының негізгі үш *массив* параметрлерінің метасипаттамасы – көркем мәтіндердің; көркем емес мәтіндердің; драмалық шығармалардың. Мысалы, ОТҰК-да көркем әдебиет мәтіндері үшін мына деректер көрсетілген:

1. мәтін жанры: жанрға жатпайтын проза, өмірбаяндық проза, детектив, балалар әдебиеті, тарихи проза, криминал әдебиет, оқиғалар, фантастика (қияли шығарма), әзіл-сықақ (юмор, сатира);

2. мәтін типі: өмірбаяндық проза, анекдот, ассоциативтік проза, боевик, детектив, очерк, әдеби хат, повесть, нақыл (притча), пьеса, әңгіме, роман, ертегі, триллер, эпопея, эссе және т.б.;

3. мәтін хронотопы: мәтінде сипатталатын оқиғалардың шамамен алғандағы орны мен мезгілі.

Нақты алғанда, мыналар ұсынылады: көне Шығыс; XVII ғасыр Ресей; XIX ғасыр Ресей; Ресей/СССР: толығымен кеңес кезеңі; Ресей, кеңес кезеңі – Германия 1920-1940-шы жылдар; Ресей/СССР – Еуропа 1960-1980-шы жылдар; Ресей/СССР: қайтақұру; Ресей/СССР: кеңестік және одан кейінгі кезең; Америка: қазіргі өмір; Израиль: қазіргі өмір; Орта Азия: қазіргі өмір; ирреалды әлем және т.б. Сонымен бірге «хронотоп анықталмаған» деген тэг те кездесуі мүмкін.

ОТҰК-дағы имплициттік метабелгіленімге жататындар:

1. «мәтін-стиль», мұндай жағдайда академиялық, ғылыми-көпшілікке, ресми-іскерлік, бейтарап, төмендетілген (сниженный), өрескел қарапайым тілдің элементтерімен және жаргон, архаизм, жекеше-авторлық, диалекті және т.б. (барлығы 21) бөлініп қарастырылады;

2. тыңдаушылар – жасы;

3. тыңдаушылар – білім деңгейі;

4. тыңдаушылар – өлшем (размер)

(толығырақ: http://ruscorpora.ru/corpora-parameter.html)

Қайсыбір мәтін болмасын оның авторы болады. Олар:

а) мәтіннің нақты авторы болған жағдайда оның аты-жөні толық көрсетіледі; ә) мәтін авторлары бірнешеу болған жағдайда ұжымдық авторлардың аты-жөні беріледі. Олар мәселен, ұжымдық монографиялар, бірлесіп жазылған мақалалар т.б.; б) жалпылама автор, мұндай мәтіндер жеке адамның емес, ұжымның, мекеменің атынан кететін мәтіндер (яғни құжаттар, хаттар т.б. мәтіндері); в) кейбір мәтіндер авторлары белгісіз де болуы мүмкін. Бұл әсіресе газет-журналдар мәтіндерінде көп кездеседі. Мұндай мәтіндер авторлары кейде шартты есімдермен де көрсетіледі. Мысалы, орыс тілінде «Иван Иванов», «Аноним», қазақ тілінде де мұндай шартты атаулар кездеседі. Мұндай авторы анық емес мәтіндердің метабелгіленімдерінде автор деген ұяшық толтырылмай бос қалдырылады.

Метабелгіленімдерде кейде авторлардың жынысына қатысты да ақпарат беріледі. Автордың әйел адам екендігі немесе ер адам екендігі немесе жынысы анық көрсетілмеуі де мүмкін. Әдетте автордың жынысы мәтін авторы біреу болған жағдайда көрсетіледі, ал ұжымдық мәтіндерде автордың жынысы көрсетілмейді. Алайда Британ ұлттық корпусында ұжымдық мәтіннің авторлары әртүрлі жынысты болған жағдайда «mixed» деген белгіленім қойылады.

Метабелгіленімнің авторға қатысты тағы бір түрі – автордың жас ерекшелігінің де көрсетілуі. Кейбір корпустарда автордың шығарманы жазған кездегі жас шамасы көрсетілсе (Британ, Чех), кейбір корпустарда автордың туған жылы, күні туралы

нақты, дәл мәліметтер беріледі немесе шамамен көрсетіледі (Орыс тілінің ұлттық корпусы). Ал автордың жасын анықтау қиын болған жағдайда «белгісіз» екендігі туралы белгі қойылады. Ал ұжымдық, жалпылама, белгісіз авторлар болған жағдайда жас ерекшелігі берілмейді. Ал кейде күнделік, жеке хаттар сияқты жеке басқа тән мәтіндер авторлары белгілі болғанмен, олардың аты-жөндері берілмей, шартты атпен беріліп, бірақ жынысы мен жасы көрсетіле береді.

Башқұрт тілінің зерттеушісі З.А.Сиразитдинов автордың, информанттың ұлтын да көрсеткен (Сиразитдинов, 2013).

Корпусқа енгізілген *мәтіннің атауы* да – негізгі метабелгіленімдердің бірі. Корпусқа енгізілген мәтіндердің атауының, яғни тақырыптарының бәрі болмауы мүмкін. Егер мәтінде тақырыптар атауы берілсе, олар метабелгіленімдер жүйесіне салынады, ал тақырыптары берілмеген мәтіндердің атаулары көрсетілмейді. Бұлар әдетте газеттер мен журналдардағы бір рубрика ішінде берілетін қысқа мәтіндер, демек, корпусқа салынған мәтіндердің барлығы да табиғи тіл қолданысын сипаттайтындықтан, тақырыбы жоқ болса да алына береді, бірақ метабелгіленімдер жүйесінде көрсетілмейді.

Мәтін туралы метабелгіленімдердің бірі – *мәтіннің жазылу уақыты*. Әдетте мұндай белгіленімдер автордың шығарманы жазу барысында мәтіннің соңында қалдырған мәліметтерінен алынады. Көбінесе мәтіннің жазылу уақыты библиографиялық, өмірбаяндық зерттеулерден анықталады. Ал мәтіннің жазылу уақыты туралы нақты ақпарат болмаған жағдайда оның мерзімі 5-10 жылы аралағында шамамен алынады. Кейде мәтіндердің жазылу уақыты туралы нақты мәлімет болмаған жағдайда корпусқа салынған мерзімі алынады.

Метабелгіленімдердің бірі – *корпусқа енгізілген мәтіндердің әрбіріндегі сөзқолданыс саны туралы ақпарат*. Корпусқа мәтін енгізуде стильдер арасалмағы негізінен теңгерімді болғанмен, кейбір жанрларға қатысты мәтіндер көлемі әртүрлі болып келеді. Мысалы хабарландырулар, құттықтаулар, жаңалықтар т.б. өте қысқа, олардағы сөзқолданыс саны да он шақты сөзден тұруы мүмкін. Сондықтан корпусқа енгізілген метабелгіленімдердегі мәтіндердегі сөзқолданыс саны оннан он мыңдаған сөзқолданысқа жетеді. Сөзқолданыс саны өте үлкен болып келетіндері – көбінесе корпусқа тұтастай енгізілген романдар немесе монографиялар сияқты көлемді еңбектер. Бірақ кейде мұндай шығармалар мен монографиялардың бәрі алынбай, белгілі бір тарауы ғана алынуы да кездеседі. Мәтіндердегі сөзқолданыс саны оларға берілген метабелгіленімдер жүйесінде көрсетіледі. Сөзқолданыс санын мәтінді енгізу барысында арнайы компьютерлік бағдарламалар арқылы анықтайды. Кейбір корпустарда мәтіндердегі сөйлемдер саны туралы да ақпараттар берілген. Мысалы, Браун корпусы.

Метабелгіленімдердің енді бір түрі мәтіннің *қолданылу саласы* болып табылады, ол мәтіннің ең жалпы типологиялық сипаттамасы. Мәселен, Орыс тілінің ұлттық корпусында 8 функционалды қолданыс саласы көрсетілген. Олар: *оқу-ғылыми, өндірістік-техникалық, ресми-іскери, публицистика, жарнама, діни, әдеби, тұрмыстық.*

Оқу-ғылыми қолданыс аясы ғылым мен білімнің әртүрлі салаларына жататын ғылыми және ғылыми-методикалық мазмұнды мәтіндерді біріктіріп қарайды. Ондай мәтіндерде осы қолданыс саласына сай тіл ақпарат беру қызметін атқарады.

Өндірістік-техникалық қолданыс аясы техникалық құрылғыларды және өндіріс-

тік үдерістерді сипаттайтын мәтіндердің қолданыс аясы болып табылады. Ол бір жағы оқу-ғылыми саламен шектессе, екінші жағынан іскери саламен де байланысты. Өндірістік-техникалық қолданыс саласында тіл ақпараттық және ықпал ету (нұсқаулықтар) қызметін атқарады.

Ресми-іскери салаға өндірістік, ауылшаруашылық, заңнамалық т.б. істер бойынша мемлекет пен жеке адамдар арасындағы, ұйымдар мен басқа мемлекеттер арасындағы, сондай-ақ ұйымаралық немесе ұйым ішіндегі немесе ұйым мүшелерінің арасындағы т.б. қатынастар сипатталатын мәтіндер жатады. Бұл салаға қатысты мәтіндер ықпал ету және ақпараттық қызмет атқарады.

Публицистика саласы халықты ақпараттандыру және саясат, экономика, өнер, ғылым, ахлақ және т.б. салалардағы қоғамдық маңызды мәселелер бойынша қоғамдық пікірлерді қалыптастастыруды мақсат ететін мәтіндерді біріктіреді. Бұл салада тіл ақпараттық, ықпал ету және біршама эстетикалық та функция атқарады.

Жарнама саласында негізінен материалдық қажеттілікті қалыптастыруға бағытталған мәтіндер алынады. Ондай мәтіндердің мақсаты – қандай да бір тауарды сату үшін тұтынушыларға ол тауар туралы жағымды ақпарат беру, сатып алуға шақыру. Мұндай мәтіндерде тіл ақпараттық және ықпал етуші қызмет атқарады.

Діни сала діни мазмұнда жазылған мәтіндерді біріктіреді. Мұндай мәтіндерде тіл ақпараттық және ықпал етуші функция атқарады.

Тұрмыстық сала адамдар арасындағы күнделікті қарым-қатынастағы бейресми түрдегі мәтіндерді негізге алады, көбінесе ауызша формада болып келеді. Тұрмыстық саладағы мәтіндердің жазбаша формалары да бар. Олар: жеке хаттар, күнделіктер, электрондық хаттар, телефондағы хаттар, құттықтаулар т.б.

Әдеби сала – автордың өмір туралы ой-пайымдаулары көрініс табатын көркем шығармашылық сала. Мұндай мәтіндерде тіл ықпал етуші ғана емес, эстетикалық та қызмет атқарады.

Мәтіндерді белгілі бір *тақырыптық топтарға* бөліп беру де метамәтіндік белгіленімдердің біріне жатады. Мысалы, *қоғамдық ғылымдар, физика, биология, саяхат, спорт, табиғат, өнер, саясат* т.б.

Алайда бұлай тақырыптық топтарға бөлу кейде шартты болып келеді. Өйткені кейбір мәтіндер бірнеше тақырып аясында, бірнеше салаға қатысты қарастырыла береді. Сондықтан белгілі бір мәтіннің тақырыптық топқа қатысын көрсеткенде, олар бір салаға ғана емес, бірнеше салаға ортақ болып та келеді. Әлемдік корпустарда (Браун корпусы, Орыс тілінің ұлттық корпусы т.б.) көркем әдебиеттерде көбінесе тақырыптық топ көрсетілмейді.

Метабелгіленімдердің енді бірі – *хронотоп*. Кейбір мәтіндерді тақырыптық топтар бойынша бөлу қиындық тудыратындықтан, кейбір корпустарда мәтіннің жазылған уақыты мен жазылған жерін көрсету де қажет болған, яғни мәтіннің *жазылған жері* мен белгілі бір *кезеңге қатысы* көрсетіледі. Мысалы, Алматы, 1998 жыл немесе Ресей, 1945-50 ж.ж. немесе Қазақстан, Кеңес өкіметі жылдары т.б.

Келесі бір метабелгіленімдердің бір түрі – *мәтін типі*. Мұнда мәтіннің белгілі бір жанрға қатысы көрсетіледі. Мысалы оқу-ғылыми сала жанрлары: мақала, монография, оқулық, реферат т.б., публицистикада: күнделік, репортаж, интервью т.б., ресми-іскери салада: заң, қарар, бұйрық, акт т.б., әдеби салада: роман, повесть, әңгіме т.б. Алайда жанр деген термин әдебиет саласында да өзіндік мәнге ие болғандықтан, онымен шатастырмас үшін корпус құрастыруда кейде оны «мәтін типі» деген терминмен де

қолданады.

Браун корпусындағы мәтіннің типологиялық сипаттамасы төмендегідей:

– Мәтін типі: әдеби, ақпараттық, ауыспалы типтер т.б.

– Жанр типі (60-қа жуық): драма, роман, музыка, философия, спорт, өндіріс т.б.

– Субжанр типі: оқулық, мақала, энциклопедия т.б.

– Мәтін типі: өлеңдер немесе проза.

Мәтін стилін көрсету арқылы мәтіннің тілдік формасы, әсіресе мәтіннің лексикалық құрамы анықталады. Олар: әдеби, әдеби емес стиль, бейтарап стиль, ресми стиль, арнайы (ғылыми) стиль т.б. Көркем прозада мынадай стильдер: бейтарап, аймақтық, қарапайым, жеке-авторлық стильдерді беруге болады.

Мәтіндерге осылайша стилистикалық сипаттама жасау корпустарда толық шешімін таппаған, сондықтан стилистикалық метабелгіленімдер жасау өзекті мәселе.

Метабелгіленімдердің келесі түрі – *аудиторияның жасын көрсету.* Мәтіннің кімге арналғанын білу мәтіннің мазмұны мен онда қолданылатын тілдік құралдарды да айқындайды. Мәтіндерге мұндай жас ерекшелігіне қарай белгіленім қою балалар әдебиетін, белгілі бір жас кезеңдеріне арналған оқулықтарды табуға мүмкіндік береді. Балалар әдебиеті әдетте 1-10 жас, жасөспірімдік кезең 11-17, жастар әдебиеті 18-34 жас аралығы болып келеді. Немесе тек үлкен адамдарға арналғанын көрсететін белгіленім қойылады. Кейде мәтіндер жас ерекшелігіне қатысты көрсетілмейді, яғни аудиториясы бейтарап болып келеді.

Метабелгіленімдердің келесі түрі – мәтіннің кімге арналғанын, бұл жерде *аудиторияның білім дәрежесіне* қарайғы ерекшелігін көрсету. Бұл жерде мәтін аудиториясының жалпы білімі немесе арнайы білімі, сондай-ақ жоғары білімі немесе білімінің төмендігі сияқты сипаттары негізге алынады. Себебі арнайы кәсіби салада жазылған мәтіндердің өзіне тән терминологиясы болады. Ешқандай кәсиби білімсіз жалпыға ортақ мәтіндер де болады.

Сондықтан аудиторияның білім дәрежесі ескеріліп, пайдаланушы ортаны көрсететін метабелгіленімдер қоюға болады. *а) жоғары білімді ортаға арналған, ә) кәсіби білімді қажет ететін, б) кәсіби білімі жоқ, жоғары білімі жоқ ортаға арналған, в) бейтарап орта мәтіндері.*

Метабелгіленімдердің келесі түрі – *мәтіндерді пайдаланушы ортаның саны, көлемі*. Кейбір мәтіндер жалпы көпшілікке арналып, яғни мыңдаған, миллиондаған адамға арналса, отыз шақты адамнан тұратын топқа немесе бір ғана адамға арналған болуы да мүмкін. Жалпы көпшілікке арналған мәтіндер көбінесе баспа беттеріне шыққан мәтіндер, электронды қарым-қатынасқа арналған мәтіндер болса, топқа арналған мәтіндер оқу лекциялары, кеңсе құжаттары т.б., ал жеке аудиторияға арналған мәтіндер көбінесе жеке хаттар болып келеді.

Метабелгіленімдердің келесі түрі – мәтін алынған *дереккөздер* болып табылады. Мәтін жинаудың түрлі әдіс-тәсілдері бар. Электронды кітапханалардан, интернетке салынған сайттардан, газет-журналдар, кітаптар шығарылатын баспалардан, жеке адамдардан алуға болады немесе қолдан сканер жасалады немесе қолдан теріледі. Жарық көрмеген мәтіндер қолжазба ретінде көрсетіледі. Интернеттен алынған мәтіндерде сайт аты беріледі. Жарық көрген газет-журналдардың шыққан уақыты алынады.

Мәтіндер жарық көрген баспа атын да көрсету – метабелгіленімнің бір түрі. Бұл әсіресе кітаптарға арналған.

Метабелгінімнің келесі түрі – *кітаптың немесе газет-журналдардың шыққан, жарық көрген жылын көрсету*.

Корпусқа енгізілген мәтіннің корпусқа енгізуге дейін қандай формада (электронды, кітап, газет-журнал, іскери құжат т.б.) болғанын көрсету де қажет.

Метабелгіленімдердің тағы бір белгіленімі есебінде *қосымша ақпараттарды көрсетуге* болады. Олар мәселен: а) Мәтіннің электронды нұсқасының сапасы жайлы ақпарат; ә) Подкорпустардың шартты атаулары туралы ақпарат (омонимдері ажыратылған немесе ажыратылмаған корпус, ауызша корпус, жергілікті корпус т.б.; б) Корпус туралы комментарийлер (мұнда корпус туралы қосымша ақпарттар); в) Корпусты құрастыруда мәтіндер алуға көмектескен, атсалысқан ұйымдар туралы ақпарат т.б.); г) Корпусты құрастырушылар мен жауапты жетекшілер туралы ақпарат т.б.

Башқұрт тілінің корпусын құрастырушылар мынадай метабелгіленімдер енгізілгендігін көрсетеді. Олар:

Мәтіннің паспорты (барлық мәтіндерге арналған):

а) мәтін авторы;

ә) мәтін атауы;

б) мәтін көлемі (сөзқолданыс саны);

г) мәтіннің жазылған уақыты.

Көркем мәтіндер үшін а) мәтін типі (повесть, әңгіме, роман, ертегі, триллер, эпопея, эссе т.б.)

Публицистикалық мәтіндер үшін:

а) *мәтін типі (журнал, газет)*;

ә) *мәтін тақырыбы, саласы:*

- саяси және әлеуметтік өмір;
- философия;
- экономика;
- ауылшаруашылық;
- өнер, мәдениет, әдебиет;
- ғылым және техника;
- білім;
- табиғат және саяхат;
- күнделікті өмір;
- спорт;
- дін;
- психология;
- медицина;
- сұлулық және денсаулық т.б.

б) мәтін жанры:

- интервью, сұхбат;
- мақала, очерк, репортаж, шолу;
- кеңестер;
- хаттар;
- құттықтаулар;
- көркем жанрлар; рецензия т.б.

Сонымен, жоғарыда метабелгіленімдердің түрлері туралы қысқаша сипаттама бердік. Бұлар алдағы уақытта Қазақ тілінің ұлттық корпусын құрастыру барысында

негізге, басшылыққа алатын мәселелер болып табылады. Корпус жасауда метабелгіленімдерді енгізу қажеттігі дау тудырмайды. Олар корпусты пайдаланушыға, әсіресе тілші-зерттеушілерге кез келген қажетті ақпаратын тез әрі оңай тауып алуға мүмкіндік береді. Сондықтан метабелгіленімдер енгізудің әлемдік тәжірибесін енгізу үшін оның теориялық мәселелерін айқындап алу қажет.

Мәтіндерге метасипаттау жүргізуге мүмкіндік беретін компьютерлік бірнеше лингвистикалық бағдарламалар бар. Ол – ең алдымен, Systematic Coder және UAM Corpus. Бағдарламалардың деректеріне еркін қол жеткізуге, тегін пайдалануға болады және корпустық лингвистикада қабылданған стандарттарға сай келеді. Осы бағдарламаның негізінде әртүрлі тілдердің пайдалануға берілген корпустарында мәтіндердің метасипаттау архитектурасы жасалған.

Енді метабелгіленім қоюдың бағдарламалық жолдары төменде сөз болады. Бұл бағдарлама қазақ тілінің ұлттық корпусының бес миллион сөзқолданысты қамтитын мәтіндері бойынша қойылған метабелгіленімнің бастапқы нұсқасы бойынша баяндалады.

Қазақ тілінің Ұлттық корпусын арнайы құрастырылған компьютерлік бағдарламалармен қамтамасыз ету жұмысы VisualStudio 2010 аспаптар ортада C# бағдарламалық тілінде жүргізілді. Ал морфологиялық белгіленім жүргізілген мәтіндер корпусын, сөйлемдерді, сөздерді MSSQLServer 2008 атты жадында сақтау қорында жүзеге асты.

Мәтіндер корпусы дегеніміз – әрбір автордың немесе стильдердің (жанрлардың) MSWORD форматындағы мәтіндер топтамасы. Әрбір шығарманың басында метабелгіленім орын алады, мысалы:

<Жанр> Проза<\Жанр>

<Автор>Әбіш Кекілбаев<\Автор>

< Тақырыбы>Аңыздың ақыры<\ Тақырыбы >

< Тақырыпша > Қызыл алма<\ Тақырыпша >

<Жылы>1983<\Жылы>

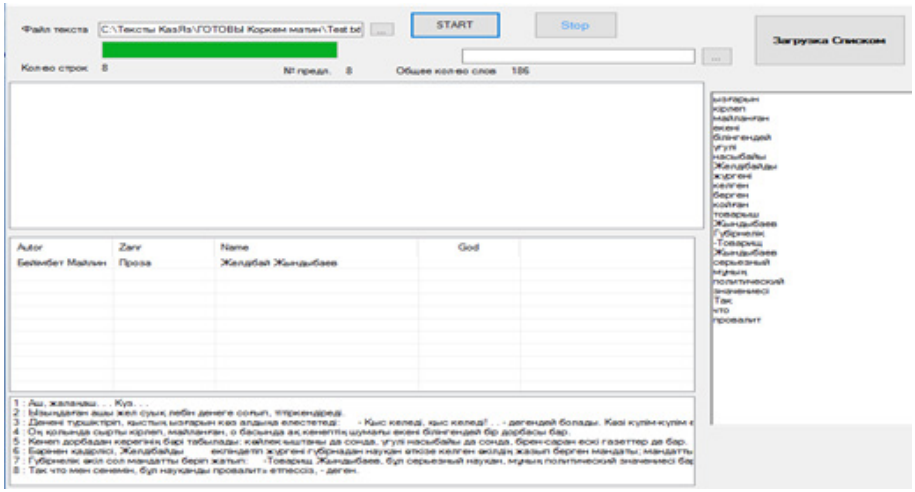
Тегтер түріндегі метабелгіленімдер бойынша шығарма жанры, шығарма авторы, шығарма аты, тақырыпша аты және баспадан шыққан жылы.

Мәтіндер файлдарын, олардың тізімі бойынша немесе жекелік сипатын автоматты түрде өңдеудің компьютерлік бағдарламасы C# тілінде жазылды.

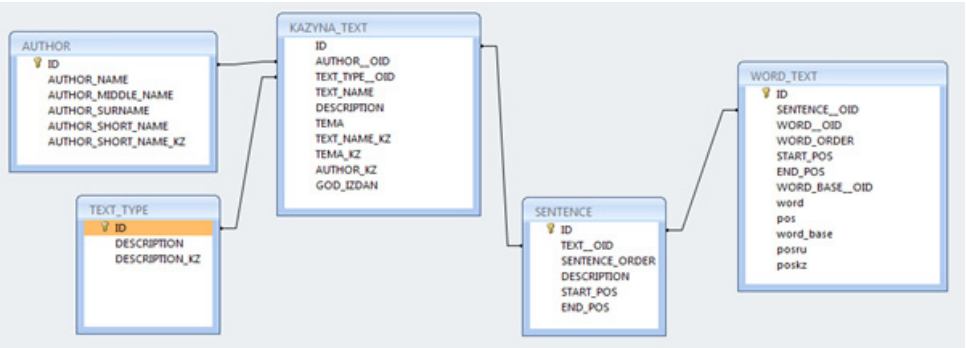
Корпус жадына енетін мәтін рет-ретімен келесі аталған өңдеу модульдерінен өтеді:

- Метабелгіленім тегтерін өңдеу модулі;
- Мәтін ішінен сөйлемдерді жекелеп бөліп алу модулі;
- Сөйлемдер ішінен сөздерді жекелеп бөліп алу модулі;
- Морфологиялық талдау (анализдеу) модулі;
- Деректер базасындағы өңделген ақпаратты корпуста сақтау модулі.

Аталған бағдарламаның мақсаты – ақпаратты іздестіру жүйесінің деректер базасын (қорын) қалыптастыру. Мысалы, корпус мәтіні ішінен кез келген сөз бойынша және қажетті деген шығарма авторы бойынша осы сөз орын алатын барлық сөйлемдерді іздеп табады. Осы мақсатқа жету үшін әрбір модуль әрекетінің нәтижесі сақталатын деректер базасын жобалау қажет (1-сурет).



1-сурет – Метабелгіленімдер енгізілген файлдарды өңдеу бағдарламасы



2-сурет – Деректер базасының сызбасы

Тегтерді өңдеу модулінің әрекеті нәтижесінде осы модульдің шыға берісінде шығарманың авторы, шығарманың аты және баспадан шыққан жылы жайлы ақпарат аламыз. Бұл мәліметтер үш түрлі кестелерде: ***Author, Kazyna_Text, Text_Type*** сақталады.

1-кесте – *Author* кестенің құрылымы

Өріс, түрі	Өріс сипаты
Id	Идентификатордың кілті
AUTHOR_NAME	Автордың аты
AUTHOR_MIDDLE_NAME	Фамилиясы
AUTHOR_SURNAME	Әкесінің аты
AUTHOR_SHORT_NAME	Автордың фамилиясы орыс. инициалымен
AUTHOR_SHORT_NAME_KZ	Автордың фамилиясы қаз. инициалымен

Авторлар кестесінен негізгі өріс Kazyna_Text кестесінің AUTHOR__OID өрісімен байланысты. Kazyna_Text кестесінде шығарма аты, шығарманың тақырыбы, жанрдың аты сақталады. Сонымен, арнайы кілт бойынша кез келген автордың барлық шығармаларын тауып алуға болады.

2-кесте – Kazyna_Text кестесінің құрылымы

Өріс, түрі	Өріс сипаты
Id	Идентификатордың кілті
AUTHOR__OID	Авторға сілтеме
TEXT_TYPE__OID	Шығарма жанры
TEXT_NAME	Шығарма аты /орыс.
DESCRIPTION	Сипаттама
TEMA	Шығарма тақырыбы
TEXT_NAME_KZ	Шығарма аты /қаз.
TEMA_KZ	Шығарма тақырыбы /қаз.
GOD_IZDAN	Баспадан шығу жылы

Kazyna_Text кестесінде TEXT_TYPE__OID шығармасының жанрына қатысты сілтеме бар. Бұл кілт 6 суретке сәйкес ID өрісі бойынша Text_Type кестесімен байланысты. Text_Type кестесінде шығармалардың барлық жанрлары сақталады.

3-кесте –Text_Type кестесінің құрылымы

Өріс, түрі	Өріс сипаты
Id	Идентификатордың кілті
DESCRIPTION	Шығарма жанрының аты

Мәтіннен сөйлемдерді бөліп алу модулінде белгілі символдарға аяқталатын (!, ?. , ,) барлық сөйлемдер орын алады. Сөйлем соңын бейнелейтін осы символдарға дейінгі барлық сөздер толық сөйлем деп саналады да Sentence кестесіне жазылады.

4-кесте – Sentence кестесінің құрылымы

Өріс, түрі	Өріс сипаты
Id	Идентификатордың кілті
TEXT__OID	Шығарма атына сілтеме
SENTENCE_ORDER	Сөйлемнің рет санының номері
DESCRIPTION	Сөйлем мәтіні
START_POS	Мәтіндегі бастапқы позиция
END_POS	Мәтіндегі соңғы позиция

Sentence кестесінде TEXT__OID өрісі бойынша Kazyna_Text кестесіне сілтеме бар. Мұндай байланыс кез келген таңдалып алынған шығарма авторы бойынша барлық сөйлемдерді тауып алуға мүмкіндік жасайды. Морфологиялық талдаудың соңғы модулінде сөйлемдердің барлық сөздерін тауып алып, оларды морфологиялық өңдеуден өткізіп, нәтижесінде сөйлемнің әр сөзіне морфологиялық белгіленім тіркеліп жазылады. Бұл модуль әрекетінің соңында біздер морфологиялық белгіленімнен өткен сөздер тізімін және сәйкес келетін мәтіндегі сөйлемнің реттік номері тіркеледі. Аталған ақпараттың барлығы да Word_Text кестесінде сақталады (5-кесте).

5-кесте – Word_Text кестесінің құрылымы

Өріс, түрі	Өріс сипаты
Id	Идентификатордың кілті
SENTENCE__OID	Сөйлемдер кестесіне сілтеме
START_POS	Сөйлемдегі бастапқы позиция
END_POS	Сөйлемдегі соңғы позиция
Word	Сөзформа

Pos	Морфологиялық белгіленім
word_base	Негіз сөз
Posru	Морфологиялық белгіленімнің қысқаша формасы /рус.
Poskz	Морфологиялық белгіленімнің қысқаша формасы /каз.

Word_Text кестесінде Sentence кестесіне сілтеме бар. Сонымен, әрбір сөз бойынша оның мәтіндегі сөйлемдермен байланысы бар екендігін, яғни шығарма аты мен шығарма авторы бойынша оларды қадағалап отыруға болады.

Қорытынды

Мұндай ғылыми-зерттеу жұмыстарына қажетті корпустан алынатын мәліметтер оған енгізілген белгіленімдердің берілуі, әсіресе тілші-мамандар, тілді зерттеушілер үшін аса қажет. Корпустағы ең маңызды рөл атқаратын лингвистикалық белгіленімдердің сапалылығы оны қолданушылардың жұмыс істеу, ғылыми материалдар жинау ісін жеделдетеді әрі мол тілдік ақпараттармен қамтамасыз етеді.

Демек, корпустың қандай түрін алсақ та, қандай мәтінмен жұмыс жасасақ та ең алдымен жинақталған мәтін туралы ақпарат беретін метамәтіндік параметрлер компьютерлік базаға салынуы қажет. Сондықтан корпус құрастыруда метабелгіленімдер (шығарма атауы, авторы, жылы, шыққан орны, стилі, жанры, көлемі т.б.) жасаудың әдіс-тәсілдерін, түрлерін, компьютерлік бағдарламаларын т.б. зерттеудің маңызы зор. Қазақ тілінің ұлттық корпусы тиімді құрастырылып, өз дәрежесінде қолданысқа енгізілуі үшін әлемдік тәжірибедегі метабелгіленімдер қоюдың ортақ ұстанымдары мен әдіс-тәсілдерін зерттеу қажет. Ал бұл жетістік қазіргі дамыған еліміздің ғылыми-зерттеушілік әлеуетін, ғылым мен білімді жан-жақты дамытатын бірден-бір күш екендігі сөзсіз.

Information about authors:

Zhanabekova Ayman Abdildakyzy – doctor of philological Sciences, A. Baitursynov Institute of Linguistics, aiman_miras@mail.ru 0000-0002-6169-7444;

Kozhakhmetova Aktoty Kozhakhmetovna – PhD student of the A. Baitursynov Institute of Linguistics, akony_8484@mail.ru 0000-0002-6359-9458.

ӘДЕБИЕТТЕР

Finegan E. (2004) LANGUAGE: its structure and use. – N.Y.: Harcourt Brace College Publishers. – 245 p.

Wikipedia [*электронды ресурс*] – URL: http://en.wikipedia.org/wiki/

Mc Enery T. (2001) Wilson, A. Corpus Linguistics. – Edinburgh: Edinburgh University Press,.

Рыков В.В. (2002) Корпус текстов как реализация объектно-ориентированной парадигмы // Труды Международного семинара Диалог. – М.: Наука.

Захаров В.П. (2005) Корпусная лингвистика: Учебно-метод. пособие. – СПб.

Савчук С.О. (2005) Метатекстовая разметка в Национальном корпусе русского базовые принципы и основные функции // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. – М. – С. 62-88.

Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш., Ибрагимова А.Д.

(2013) Информационные системы и базы данных башкирского языка. – Уфа: Книжная палата РБ. – 116 с.

REFERENCES

Finegan E. (2004) LANGUAGE: its structure and use. – N.Y.: Harcourt Brace College Publishers. (in English)

Wikipedia — URL: http://en.wikipedia.org/wiki/ (in Russian)

McEnery T., Wilson, A. (2001) Corpus Linguistics. – Edinburgh: Edinburgh University Press. (in English)

Rykov V.V. (2002) Corpus of texts as an implementation of the object-oriented paradigm // Proceedings of the International Seminar Dialogue-2002. – M.: Nauka. (in Russian)

Zakharov V.P. (2005) Corpus linguistics: Teaching method. allowance. – SPb. (in Russian)

Savchuk S.O. (2005) Metatext markup in the Russian National Corpus basic principles and main functions // Russian National Corpus: 2003–2005. Results and prospects. – M. – S. 62–88. (in Russian)

Sirazitdinov Z.A., Buskunbaeva L.A., Ishmukhametova A.Sh., Ibragimova A.D. (2013) Information systems and databases of the Bashkir language. – Ufa: Book Chamber of the Republic of Belarus.- 116 p. (in Russian)