МРНТИ 16731700

А.Қ.Жұбанов

А.Байтұрсынұлы атындағы Тіл білімі институтының
бас ғылыми қызметкері, филология ғылымдарының докторы, профессор

ҚАЗАҚ ТІЛІНІҢ ҰЛТТЫҚ КОРПУСЫНДАҒЫ ІЗДЕСТІРУ АППАРАТЫНЫҢ ЕҢ ҚҰНДЫ БӨЛІГІ – МӘТІНДЕРГЕ ТІРКЕЛІП БЕРІЛЕТІН МЕТАБЕЛГІЛЕНІМ

Аннотация. Қазақ тілінің ұлттық корпусында қабылданған метабелгіленім жүйесі метамәтіндік ақпаратпен бірге беріледі. Бұл жерде біз метабелгіленім ұғы- мын экстралингвистикалық сипаттағы немесе сыртқы аннотациялаудың және мәтінге қатысты техникалық жұмыстардың (яғни қызметтік) белгілер жүйесі деп түсінуіміз қажет. Корпустық лингвистикаға қатысты әдебиеттерді зерделей келе, белгіленімнің бірнеше түрі болатындығы айқындалды. Солардың ішінен экстра- лингвистикалық белгіленім немесе метабелгіленім (метаразметка) деп аталатын, мәтіндік деректер жайлы мағлұмат беретін белгіленім түрін де қарастырдық.

Тірек сөздер: метабелгіленім, экстралингвистикалық белгіленім, метамәтіндік ақпарат, көркем проза, поэзия, ғылыми-гуманитарлық стиль, ғылыми-техникалық стиль, публицистикалық стиль.

А.К.Жубанов

главный научный сотрудник Института языкознания им. А. Байтурсынова,
доктор филологических наук, профессор

САМАЯ ЦЕННАЯ ЧАСТЬ ПОИСКОВОГО АППАРАТА В НАЦИОНАЛЬНОМ КОРПУСЕ КАЗАХСКОГО ЯЗЫКА - МЕТАРАЗМЕТКА, КОТОРАЯ ХАРАКТЕРИЗУЕТ ТЕКСТ В ЦЕЛОМ

Аннотация. Система метаразметок, принятая в Национальном корпусе казахского языка, сопровождается метаразмеченной информацией. Здесь мы должны понимать метаразмеченную систему признаков экстралингвистического характера или внешнего аннотирования и технических работ (т. е. служебных), относящихся к тексту. На основе изучения литературы, касающейся корпусной лингвистики, было установлено, что существует несколько видов разметок. Среди них мы рассмотрели так называемую экстралингвистическую разметку или метаязык (метаразметка), который дает представление о текстовых данных.

Ключевые слова: метаразметка, экстралингвистическая разметка, метатекстовая информация, проза, поэзия, научно-гуманитарный стиль, научно- технический стиль, публицистический стиль.

A. K. Zhubanov

Chief Researcher of the Institute of Linguistics named after A. Baitursynov,
Doctor of Philology, Professor

THE MOST VALUABLE PART OF THE SEARCH ENGINE IN THE NATIONAL CORPUS OF THE KAZAKH LANGUAGE IS META -MARKUP, WHICH CHARACTERIZES THE TEXT AS A WHOLE

Annotation. The system of meta-markups adopted in the National Corpus of the Kazakh Language is accompanied by meta-marked information. Here we should understand the meta-marked system of features of extralinguistic nature or external annotation and technical works (i.e., service works) related to the text. On the basis of the study of the literature on corpus linguistics, it was found that there are several types of markup. Among them, we have considered the so-called extralinguistic markup or metalanguage (meta-markup), which gives an idea of text data.

Keywords: meta-markup, extralinguistic markup, metatext information, prose, poetry, scientific and humanitarian style, scientific and technical style, journalistic style.

Біздің «Қазақ тілінің ұлттық корпусына метамәтіндік белгіленім енгізудің ұстанымдары» атты зерттеу тақырыбымызға қатысты әлемдік тіл біліміндегі тәжірибелермен таныса келе, мәтіндерге метабелгіленім енгізудің ең ұтымды деген әдіс-тәсілдері анықталды. Нәтижесінде қазақ мәтіндерінің әртүрлі жанрлары мен стильдеріне қатысты шығарма мәтіндері таңдалып алынды:

1) көркем проза;
2) поэзия;
3) ғылыми-гуманитарлық стиль;
4) ғылыми-техникалық стиль;
5) публицистикалық стиль (газет пен журналдардағы мақалалар);
6) драмалық шығармалар мәтіндері және т.б.

Таңдалып алынған стильдердегі мәтіндерге метабелгіленім енгізу мақсаты ең алдымен корпусты пайдаланушының міндеттерімен сай келуі қажет. Сонымен бірге, пайдаланушыға енгізілген белгіленімдерде кейбір қателердің де кездесуі мүмкін екені жайлы ескертіліп қойылғанын да жөн санаймыз. Солай бола тұра, мәтінге жүргізілген метамәтіндік белгіленім әдісі аса пайдалы құрал ретінде қолданыс табуда.

Белгіленім негізіне көпшілік мақұлдаған және, мүмкіндігінше, теориялық жағынан алғанда бейтарап түрдегі лингвистикалық қағидалар алынуы қажет. Сонымен бірге, белгіленімнің ешқайсысын да стандарт деп санауға болмайтынын ескеру қажет.

Корпустық лингвистикаға қатысты әдебиеттерді зерделей келе, белгіленімнің бірнеше түрі болатындығы айқындалды. Солардың ішінен **экстралингвистика-лық белгіленім** немесе **метабелгіленім (метаразметка)** деп аталатын, мәтіндік деректер жайлы мағлұмат беретін белгіленім түрін қарастырайық.

Метабелгіленімді, шартты түрде, сыртқы, құрылымдық және техникалық деп бөліп-бөліп қарастыруға болады. Мұндағы сыртқы белгіленім бойынша автор жайлы және мәтін туралы мәліметтер беріледі (автор, шығарма аты, басылымның жылы мен орны, жанры мен тақырыбы). Құрылымдық белгіленім бойынша тараулар, абзацтар, сөйлемдер мен сөзформалар маркаланады (белгіленеді). Техникалық белгіленім мәтінді өңдеу мезгілін, орындаушыларды және электрондық нұсқаның дереккөзінің кодталуын белгілейді. Метабелгіленім тілдің өмір сүру жағдаятын зерттеу үшін, тілдегі өзара ішкі байланыстарды анықтауға және тілдегі дербес ішкі тілдік жиындарды зерделеуге қажет.

Қазіргі кезде метабелгіленімдерді стандарттау мәселесіне көп көңіл бөлінуде:

– TEI (Text Encoding Initiative) жобасы,
– EAGLES (Expert Advisory Group on Language Engineering Standards) ұсынысы,
– CES (Corpus Encoding Standard) стандарты,

– XCES (Corpus Encoding Standard for XML) стандарты,
– ISLE (International Standards for Language Engineering) жобасы,
– CDIF (Corpus Document Interchange Format, BNC) стандарты.

Метамәтіндік ақпаратты енгізу мәтінді толығымен сипаттайды. Мұндай ақпараттың қазақ тілінің ұлттық корпусында орын алуы зерттеушіге мәтіндердің миллиондаған ауқымындағы тілдік фактілер мен құбылыстарды іздестірудің шеңберін сызуға мүмкіндік туғызады.

Қазақ тілінің ұлттық корпусында қабылданған метабелгіленім жүйесі метамәтіндік ақпаратпен бірге беріледі. Бұл жерде біз метабелгіленім ұғымын экстралингвистикалық сипаттағы немесе сыртқы аннотациялаудың және мәтінге қатысты техникалық жұмыстардың (яғни қызметтік) белгілер жүйесі деп түсінуіміз қажет.

Қазіргі кезде мәтіндердің метасипаттамасын жүргізуге мүмкіндік беретін бірнеше лингвистикалық бағдарламалар бар. Ең алдымен оған жататын – Systematic Coder және UAM Corpus. Бұл аталған бағдарламалар ашық қолжетімді сипатта. Бұлардың бәрі корпустық лингвистикада қабылданған стандарттарға сәйкес келеді. Осылардың негізінде қазақ тіліндегі мәтіндерге метасипаттама берудің архитектурасы жасалды деуге болады.

Қазақ мәтіндерін өңдеу барысында қазақ тілі корпусындағы мәтіндердің метамәтіндік ақпаратын MS Access 2007 құрастырған деректер базасында (қорында) сақтаудың қажеттігі жайлы шешім шығарылады деген ойдамыз. Аталған деректер базасы өзара байланыстағы үш кестеден тұрады: «Authors», «Books» және «Texts».

«Authors» кестесі авторлардың сипаттамаларын қамтиды. Автордың фамилиясы және аты-тегі үш тілде тіркеледі (қазақша, орысша, ағылшынша). Егер автордың аты-тегі белгісіз болса, кестеде «авторы белгісіз» және бірнеше автор болса, кестеде «ұжымдық автор» деп тіркеледі.

Аталған кестеде автор жайлы толық сипаттама да берілген:

– Өмір сүрген жылдары (туған жылы және, егер ол мүмкін болса, қайтыс болған жылы);
– жынысы (ер немесе әйел адам);
– туған жері;
– алғашқы тілі (қарым-қатынас пен таным-білім алуда алғашқы қолданылған тіл);
– диалект;
– басқа тілдерді меңгеруі;
– білім деңгейі;
– жұмыс түрі.

«Books» деп аталған екінші кестеде, негізінен, қызметтік сипаттағы кітаптар сипатталады. Әрбір кітап өзінің бірегей нөміріне ие (өріс түрі: тіркеуіш) және осы жерде редакторлар да белгіленеді, әдетте, олар – ақын-жазушылар. Аудармашы (Translator) деп аталатын жеке өріс параллель (қатар) подкорпустар (корпусшалар) үшін қажет. Кітаптың шығу мәліметтері библиографиялық сипаттамада тіркеледі және олардың әрбір элементін жеке өрістерде белгілемей, барлық библиографиялық сипаттамаларды бір ғана өрісте біріктіріп сақтау қажет деп ұйғарған болатын. Әрі қарай қызметтік өрістер өз жалғасын табады:

– мәтін алынатын дереккөз (баспа, сканирлеу, электрондық көшірмесі, қолмен теру, интернет);
– орындаушылар.

Кестедегі мәтіннің аты жазылатын бағана индекстелетін өріс болып саналады және қатаң түрде қайталаулар жасалмауы керек: базада бұрын жазылған деректі қайталап жазуға болмайтындығы жайында база жүйесі ескертіп отырады.

Егер мәтін аударма әдебиетіне қатысты болса, «Автор» және «Аудармашы» өрістеріндегі (түсіп қалатын) тізімнен мәтін авторы мен аудармашы таңдалып алынады.

Келесі өрістер мәтін атрибуттерінің сипаттамасын өз бойында сақтайды:

– тіл формасы: жазба немесе ауызша;
– графикалық жүйе;
– тіл типі (түрі): көркем, ресми-іскерлік, ғылыми, ауызша сөйлеу тілі және т.б.;
– көптеген элементтерді өзінде сақтайтын мәтін жанры;
– шығу мәліметтері;
– беттер;
– орындаушы;
– тіркеу күні;
– түсіндірме.

Деректер базасы метаақпаратты қайта түсіндіру тәсілі деуге болады. Кестелерді толтыру барысында мынадай жағдайды ескергенді жөн санаймыз, ол поэтикалық стильдер мәтіндері бойынша тіл заңдылықтарын (қасиеттерін) зерттеу айтарлықтай күрделі екендігі. Оның себебі өлеңдер мәтіні көркем әдебиет мәтініне қарағанда басқаша ережелерге бағынатындығы. Өлеңдер мәтініндегі сөз метафоралық немесе метонимикалық мағынада қолданыс табуы мүмкін. Ақындар тілін құру мәселесінің күрделілігі, оның әр авторға ғана тән қасиеттерінің орын алуы және сол себепті ақын тілінен әмбебаптық пен тәндік қасиеттерді арнайы бөліп алу аса күрделі мәселе болып саналады.

Ұлттық тілдер корпусындағы деректер базасын әртүрлі ғылыми мақсатта автономды (дербес) сипатта қолдануға мүмкіндік бар. Мысалы, құрастырылған деректер базасындағы мәтіндер тақырыптары арнайы кішігірім корпусын (подкорпусын) жасауға материал болуы мүмкін. Бұл деректер базасын бір тақырыпқа ғана сәйкес келетін материалдарды іздеп-табу үшін де қолдануға болады және ол материалдар хронологиялық пен авторлық деңгейлердегі өзгерістерді бақылап зерттеуге мүмкіндік жасайды.

Әлемдік тіл білімінде метабелгіленімді енгізудің теориялық жағдайы мен практикалық әдістері бар. Ал қазақ тіл білімінде бұл мәселе қазіргі кезде «Қазақ тілінің ұлттық корпусын» жасау мәселесіне қатысты тұңғыш рет қарастырылуда.

Қазіргі кезде лингвистикалық деректердің компьютерлік базасын құрастыру әлемдік деңгейде, дәлірек айтқанда, Америка мен Францияның, Ресей мен Белоруссияда және т.б. ғылыми орталықтарында айтарлықтай ғылыми нәтижелерге жетті деуге болады. Мысалы, ең алғашқы аса көлемді компьютерлік корпус – Браун корпусы (БК) 1960 жылы Америкада Браун университетінде жасалған болатын. Сол сияқты, осы корпусқа ұқсас Швецияның Упсал университетінде құрастырылған корпустың да маңыздылығы зор. Ресейде орыс тілінің ұлттық корпусын құрастыру 2001 жылдан бастама алады. Қазіргі кезде ұлттық корпус жасау мәселесімен Москва, Петербур, Воронеж және т.б. қалалардағы белгілі маман-ғалымдар корпус мәселесімен айналысып, қомақты нәтижелерге жеткені мәлім.

2015-2017 жж. ҚР БЖҒМ ҒК А.Байтұрсынұлы атындағы Тіл білімі институтының қазақ тілінің қолданбалы саласымен айналысатын ғалымдар **«Қазақ тілінің ұлттық корпусындағы метамәтіндік белгіленімдер ұстанымдары мен әзірлемесі»** деп аталатын гранттық жоба тақырыбына қатысты ғылыми-зерттеу жұмысын жүргізуде. Аталған тақырып бойынша қазақ тілінің ұлттық корпусына көлемі 7 млн.-дай болатын әртүрлі стильдерге қатысты мәтіндерге метамәтіндік белгіленімдер жүргізудің теориялық және практикалық мәселелерін зерттеп, қажетті нәтижелерге қол жеткізді. Алынған нәтижелерді теориялық және практикалық материалдар ретінде қазақ тілінің ұлттық корпусын құрастыру мәселесінде пайдалануға болады демекпіз.

Жоғарыда айтылғандай, қазақ тілінің ұлттық корпусы мәтіндердің электрондық түрінің әртүрлі стильдер мен жанрларының жиынтығынан тұруымен бірге олар кең түрдегі лингвистикалық және метамәтіндік ақпаратпен жабдықталады. Мәтіндердің мұндай ақпараттарға ие болуы әдеттегі интернет желісіндегі қолжетімді мәтіндер жиынтығынан ерекшелендіреді және ұлттық корпустардың ең басты айырмашылығы болып саналады. Ақпараттың егжей-тегжейлігі мен нақтылығы және сонымен бірге, мәтіндерде орын алатын кең түрде қамтылған әртүрлі тілдік деректер мен тілдік құбылыстар ұлттық корпустың бірегей лингвистикалық ресурс ретіндегі ең басты құндылығын айқындайды.

Метабелгіленімді корпус мәтіндеріне енгізу арнайы ізденіс нәтижелері негізіндегі компьютерлік бағдарламалар бойынша жүзеге асады. Зерттеу барысында мәтінге қосымша ақпараттар енгізу (тіркеу) үшін метадеректерді таңдау қағидасы зерделеніп, мәтіннен метабелгіленімдер жүйесі негізінде қажетті деген ақпаратты іздеп-табуға қатысты тиімді компьютерлік бағдарламалар құрастырылды. Корпустың іздестіру аппаратының ең құнды бөлігі деп мәтіндерге тіркеліп берілетін метабелгіленімді (немесе метасипаттаманы) айтуға болады.

Қазақ тілінің ұлттық корпусындағы мәтіндердің көлемі мен әртиптілігін ескере келе, мұндай стильдік, жанрлық, типтік деңгейдегі дифференциялау аса қажет деуге болады. Себебі, зерттеушілердің көпшілігі корпустың толық түрімен емес, зерттеу мақсатына қарай, оның тек маңызды ішкікорпусымен (подкорпусымен) ғана, мысалы, көркем әдебиет, публицистикалық, іскерлік және т.б. стильдерге қатысты зерттеулер қызықтыруы мүмкін. Сонымен бірге, метабелгіленім арқылы жүргізуге болатын қызықты мәселенің бірі – метабелгіленім параметрлерінің арасындағы (мысалы, автордың жынысы мен жасы арасындағы) және мәтіннің тілдік ерекшеліктерінің статистикалық ең анық корреляциясын (арақатынастылығын) анықтау.

Қазақ мәтіндерін ұлттық корпусқа енгізу үдерісі мынадай кезеңдерден тұрады:

1) минималды HTML- форматта мәтінге алдын ала белгіленім енгізу;

2) морфологиялық белгіленім енгізу мен омонимдерді ажырату (корпус бөлігіне);

3) метамәтіндік белгіленім;

4) Яндекс-сервер үшін шығару форматына өзгерту (түрлендіру).

Әрбір келесі кезеңде, алдыңғыға қарағанда, қосымша ақпараттың көлемі мен мазмұндылығы әр уақытта (тұрақты түрде) өсіп отырады. Бірінші кезеңде мәтінге оның формалды құрылымы жайында ақпарат енгізіледі, мәтін элементтерінің әртүрлі типтеріне, рәсімдеу параметрлеріне, арнайы символдарға белгіленім жүргізіледі. Екінші кезңде мәтінге шын мәніндегі тілдік (морфологиялық) ақпарат қосылады. Үшінші кезеңде мәтіннен тыс дайындалатын метамәтіндік атрибуттер (анықтауыштар) жинағы түріндегі мәтін «паспорты» көрініс табады. Соңғы кезеңде метамәтіндік ақпарат мәтінмен бірігіп (қосылып) және тағы да бірнеше өзгерістерге (трансформацияға) ұшырап, Яндекс-сервер арқылы жүктеледі және индекстеледі. Тек осыдан кейін ғана мәтін корпустың бөлігіне айналып, ізденіске қолжетімді болады.

Әрине, кейбір амал-әрекеттерді (омонимдерді ажырату, метабелгіленім жүргізу) автоматты түрде жүзеге асыру мүмкін бола бермейді, бірақ оларды адам қатынасы-мен жартылай автомттандыратын қолайлы орта жасауға мүмкіндік бар.

Қазіргі кездегі белгіленім жүргізетін тілдердің көпшілігі SGML/XML-ге негізделеді де белгіленімнен өткен мәтін екі параллель қабаттан тұратын ақпаратқа ие: көзге көрінетін (мәтіннің өзі) және көзге көрінбейтін (белгіленім).

Корпустық ақпаратты кодтайтын (шартты белгілер енгізу) стандарттардың ішінен айрықша абыройға ие болғандар мыналар: TEI (Text Encoding Initiative), XCES (XML

Corpus Encoding Standard), EAGLES (European Advisory Group on Language Engineering Standards).

Аталғандардың ішінен ең бір жете құрастырылғаны TEI стандартын атауға болады. Ол стандарт кез келген мәтіннің және мәтіндік ақпараттар элементтерінің ережелерін анықтай алады. Бұл жерде аталған мәтіндік ақпарат элементтеріне мыналар жатады: құрылымы, тақырыптар, тіл типтері (проза, поэзия, драма), беттері, дәйексөздер, сілтемелер, түзетулер, кестелер, формулалар, арнайы символдар, лингвистикалық аннотациялар және т.б. XCES стандарты TEI стандартының дамыған түрі және ол тек корпус мәселесіне ғана арналған және ол корпустарға тән тегтердің жиынтығын анықтайды, бірақ бұл стандарт қолданыстағы корпустарда сирек қолданылып жүргенін байқауға болады.

Іздестіру жүйесіне қатысты мәселелерді зерттей келе, оған мынадай талаптардың қойылатынын тұжырымдауға болады:

1) сөздер мен сөзтіркестердің грамматикалық, семантикалық және т.б. белгілеріне қарай, оларды корпустан іздестіру;

2) контекст пен сөздер арасындағы қашықтықты ескеру;

3) метамәтіндік ақпаратты іздестіру;

4) логикалық жалғаулықтарды, жақшаларды және контекстік операторларды қамтитын сауалдардың дамыған тілі;

5) индекстеудің тиімділігі;

6) кез келген аса күрделі сауалдарға жоғары жылдамдықпен жауап қайтару.

Қолжетімді бағдарламалық құралдардың ішінен корпус үшін іздестіру жүйесінің рөлін атқаратын келесі бағдарламалық кластарды атауға болады: XML-бағытындағы деректер базасы; толықмәтінді іздестіру жүйелері.

Іздестіру әрекеті келесі мүмкіндіктерге ие арнайы құрастырылған компьютерлік бағдарламалар арқылы жүзеге асады:

– қолданыста бар немесе алдын ала берілген мәтіндер корпустарынан сөздерді, сөзформаларды іздеп-табу;

– барлық корпус мәтіндерінен немесе белгілі бір мәтіндерден, контекстерден сөздер мен сөзформалардың қолдану статистикасын ұсыну;

– берілген сөз, сөзформа, сөзтіркестермен бірге қолданылатын сөз бен сөзтіркестің конкорданстарын анықтау;

– тілдің лексикасы мен грамматикасының статистикалық сипаттамаларын айқындау (мысалы, белгілі бір септіктегі сөздің мәтіндер корпусында қанша рет қолданылуын анықтау);

– әртүрлі мезгілдік кезеңдерге қатысы бар (мысалы, XX ғ. ортасынан кейінгі) сөзқолданыстың салыстырмалы сипаттамасын ұсыну.

Сөздердің немесе сөйлемдердің сипаттамалары (морфологиялық, семантикалық және т.б.) бойынша іздестіру кезінде сервер бұрынғы сауалдардағы сипаттамаға сәйкес келетін индекстерді ашады, сосын ашылған индекстерден өту нәтижесінде барлық қажетті деген сөзге қатысты позициялар табылады. Корпусты пайдаланушының әрекетін жеңілдету үшін арнайы сауалдар формасы қарастырылған, ол бойынша пайдаланушы айтарлықтай түсінікті түрде сұрақ беруіне мүмкіндігі бар. Жаңа формада қалыптасқан сауал іздестіру серверіне жіберіледі, ал онда арнайы *C++* тілдегі компьютерлік бағдарлама бойынша шығару модулі қалыптасады. Қойылған сауалға жауап ретіндегі қалыптасқан модуль ***xml***-форматындағы нәтиже түрінде көрініс табады, ал сосын барып оған xslt-түрлендіру қосылып, нәтижесінде пайдаланушы ***html*** форматындағы іздестіріліп отырған мәліметке ие болады.

Аталған тәсіл іздестіру әрекетін нәтижелерді дайындаудан бөліп алып қарастырады

да жүйені құру мен түрлендіруді жеңілдетеді. Зерттеу барысында байқағанымыздай, қазақ тілінің ұлттық корпусын пайдаланудың нәтижелі болуы аса көлемді мәтіндер массивінен контекстік іздестіруге арналған арнайы оңтайландырылған толықмәтінді іздестіру жүйесінің құрастырылуын қажет етеді. Мұндай жүйелердің айрықша танымал түрлері Яндекс және Google іздестіру интернет-серверлері және индекстеу мен интернет-ресурстарды іздестіруге бағытталған басқа да жүйелер. Орыс тілінің ұлттық корпусы үшін табиғи таңдау аса жоғары нәтижелігімен және ауқымдылығымен ерекшелінетін Яндекс-серверге түскені мәлім. Белгілі болғандай, Яндекс-сервер аса көлемді мәтіндер массиві бойынша және аса күрделі сауалдарға жауап алу үдерісін аса тездетеді, тіпті секунд бөлігі ішінде орындайды және корпус көлемінің ұлғаюы іздестіру жылдамдығына еш әсер етпейтіні де мәлім болып отыр.

Сөз соңында айтайық дегеніміз, қазақ тілінің ұлттық корпусын құрастырушылардың алдында тұрған негізгі мәселе, біріншіден, қазақ мәтіндерінің репрезентативті корпусын қалыптастыру, екіншіден, бұл корпус лингвистикалық белгіленім енгізілген, яғни тілдік зерттеулер жүргізу мақсатымен, ең алдымен, мәтіндерге морфологиялық белгіленім мәселесі жүзеге асырылған және қазақ тілінің барлық функционалды-стильдік қабаттары қамтылған корпус болуына барынша атсалысуды қажет етеді.

ӘДЕБИЕТТЕР ТІЗІМІ:

[1] Сичинава Д.В. К задаче создания корпусов русского языка. [электрон ресурс]. http://www.mccme.ru/ling/mitrius/article.html. (жүгіну уақыты: 2.05.2016).

[2] Демская-Кульчицкая О.М., Семеренко В.Р., Ющенко Р.А. Методы автоматической разметки текстов Национального корпуса языка // Компьютерная математика. – 2005. № 2. 6 с.

[3] Азарова И.В. Морфологическая разметка текстов на русском языке с использованием формальной грамматики AGFL. Кафедра математической лингвистики СПб.: ГУ [электрон ресурс]. //http://www.dialog-21.ru/Archive/2003/AzarovaAGFL. htm. (жүгіну уақыты: 2.05.2016).

[4] Национальный корпус русского языка. [электрон ресурс]. //http://www. ruscorpora. ru/. (жүгіну уақыты: 2.05.2016).

[5] Дарчук Н.П. Автоматизований морфологічний аналіз тексту. [электрон ресурс]. // http://linguist.univ.kiev.ua/courses_morph.htm.(жүгіну уақыты: 2.05.2016).