МРНТИ16.31.21

А.Ә.Жаңабекова[1], А.Қожахметова[2]

[1]А.Байтұрсынұлы атындағы Тіл білімі институты Қолданбалы лингвистика бөлімінің меңгерушісі, филология ғылымдарының докторы. Алматы қаласы, Қазақстан

[2]А.Байтұрсынұлы атындағы Тіл білімі институты Қолданбалы лингвистика бөлімінің лаборанты, магистр. Алматы қаласы, Қазақстан

ЖАРТЫЛАЙ АВТОМАТТЫ МЕТАМӘТІНДІК БЕЛГІЛЕНІМДЕР ЕНГІЗУ БАҒДАРЛАМАСЫНЫҢ ЖҰМЫС ІСТЕУ ТЕХНОЛОГИЯСЫ

Аннотация: А.Байтұрсынұлы атындағы Тіл білімі институты Қолданбалы лингвистика бөлімі 2009 жылдан бастап «Қазақ тілінің ұлттық корпусын» жасау мәселесімен айналысып келеді. Бұл ретте қазақ тілінің 5 стилі бойынша бірнеше миллион сөзқолданыстан тұратын мәтіндер корпусқа салынып, әртүрлі лингвистикалық және экстралингвистикалық белгіленімдер әзірлемесі дайындалды. Корпусқа енгізілетін мәтіндер туралы ең бірінші экстралингвистикалық белгіленім қою қажет. Өйткені пайдалануға ең алдымен іздеген материалының авторы, қандай мәтін екені туралы ақпарат қажет. Мұндай ақпараттарды метамәтіндік белгіленім деп атайды.

Мақалада А.Байтұрсынұлы атындағы Тіл білімі институты Қолданбалы лингвистика бөлімінде жасалған мәтіндер корпусына енгізілген 23 параметрден тұратын метамәтіндік белгіленімдер туралы баяндалады. Метамәтіндік белгіленімдердің түрлеріне арналғаннұсқаулықтар көрсетіледі. Сонымен қатар мақалада метамәтіндік белгіленімге енгізу бағдарламасының жұмыс істеу технологиясы сипатталады және корпусқа енгізілгеннен кейінгі метамәтіндік белгіленім терезесі көрсетіледі.

Тірек сөздер: корпус, аннотацияланған корпус, метамәтіндік белгіленім, терезе, іздеу.

А.А. Жанабекова[1], А. Кожахметова[2]

1заведующий отделом Прикладной лингвистики Института языкознания имени А. Байтурсынова, док.филол.н., Алматы, Казахстан

2лаборант отдела прикладной лингвистики Института языкознания имени А. Байтурсынова, магистр, Алматы, Казахстан

ТЕХНОЛОГИЯ РАБОТЫ ПРОГРАММЫ ВНЕДРЕНИЯ ПОЛУАВТОМАТИЧЕСКИХ МЕТАТЕКСТОВЫХ РАЗМЕТОК

Аннотация. Отдел прикладной лингвистики Института языкознания им. А.Байтурсынова с 2009 года занимается вопросами создания «Национального корпуса казахского языка». При этом в корпус были помещены тексты на 5 стилях казахского языка, состоящие из нескольких миллионов словоупотреблений., подготовлены различные лингвистические и экстралингвистические обозначения.

О текстах, вводимых в корпус, необходимо поставить первую экстралингвистическую разметку. Потому что в первую очередь нужна информация об авторе материала, который ищут и информация о том, что это за текст. Такую информацию называют метатекстовой разметкой.

В статье описаны метатекстовые разметки, состоящие из 23 параметров, которые внедрены в корпус текстов, созданный в отделе Прикладной лингвистики Института языкознания имени А.Байтурсынова. Указываются инструкции для различных видов метатекстовых разметок. А также в статье описывается технология работы программы внедрения метатекстовых разметок и указывается окно метаразметки после внедрения в корпус.

Ключевые слова: корпус, аннотированный корпус, метатекстовая разметка, окно, поиск

A.A. Zhanabekova[1], A. Kozhakhmetova[2]

[1]Head of the Department of Applied Linguistics, A. Baitursynyly Institute of the Linguistics, Doctor of Philology, Almaty, Kazakhstan

[2]Laboratory Assistant, Department of Applied Linguistics, A. Baitursynyly Institute of the Linguistics, Master, Almaty, Kazakhstan

TECHNOLOGY OF WORK OF THE PROGRAM OF INTRODUCTION OF SEMI-AUTOMATIC META-LAYOUT

Annotation. Department of Applied Linguistics, A. Baitursynuly Institute of the Linguistics has been dealing with the creation of the «National Corpus of the Kazakh Language» since 2009. At the same time, according to 5 styles of the Kazakh language, texts have been formed into a corpus, consisting of several million word combinations, various linguistic and extralinguistic designations have been prepared. the first extralinguistic markup, because to use, first of all, you need information about which text is the author of the searching material. Such information is called metamatic quotation.

The article considers metatext markings, consisting of 23 parameters, which are embedded in the corpus of texts created in the Department of Applied Linguistics of A. Baitursynuly Institute of the Linguistics. Instructions are provided for various types of metatext markup. Also, the article describes the technology of the program for the implementation of meta-text markup and indicates the window of meta-markup after the implementation in the corpus.

Keywords: corpus, annotated corpus, metatext markup, window, search

Корпустар атқаратын функциясына қатысты түрлі шағынтоптарға (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 сұрыптан өткізіп, олар туралы нақты ақпараттар беру қажет.

Корпусқа қойылатын белгіленім түрлерін шартты түрде лингвистикалық және одан сырт түрі деп екіге бөліп қарастыруға болады. Белгіленімнің лингвистикалықтан сырт түріне мыналар жатады:

1) мәтінді форматтау ерекшеліктерін бейнелейтін белгіленім (тақырыптар, абзацтар, бос жер және т.б.);

2) автор мен мәтінге қатысты мәліметтерді бейнелейтін белгіленім. Сонымен бірге автор жайлы мәлімет тек оның аты-жөні ғана емес, оның жасын, жынысын, өмір сүрген жылдарын және т.б. болуы мүмкін. Ал мәтін туралы мәлімет, әдетте, тақырыптан басқа оның қай тілде жазылғаны, жылы, баспа орны, аты және т.б. Мұндай ақпараттың корпуста орын алуы мәтіндер қорынан іздеу әрекетін тәптіштеп іздестіруге мүмкіндік жасайды және, сонымен бірге, олар тиісті құжатты теңестіру әрекеті үшін де қажетті құрал бола алады. 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 дерек-көз, өңделген күні, кодталған кезі т.б.) деп те бөледі.

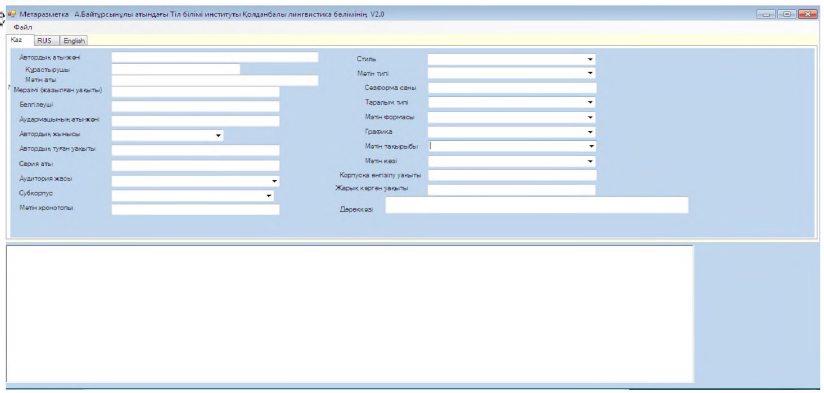
Метамәтіндік белгіленімдер корпустың қай түріне болмасын енгізілуге тиісті маңызды ақпараттар болып табылады, сонымен қатар лингвистикалық белгіленімдер қойылатын корпустар құрастыру жұмысынан бұрын, корпусқа әртүрлі стильдерден алынған мәтіндерді енгізу барысында атқарылады.

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 тарамтарам болып жіктеліп кете береді. Корпус жадына енгізілген мәтіндерді жүйелі құрылыммен беру метабелгіленімдер қоюдың теориялық және практикалық әдістәсілдерін меңгергенде ғана мүмкін болады. Сондықтан корпустың қайсібір түрінде болмасын, метабелгіленімдер қою мәселесі дұрыс жүргізілуі керек.

Метабелгіленімдер – ғылыми-зерттеу жұмыстарында белгілі бір кезеңге, стильге, авторға, тақырыпқа т.б. қатысты материалдар жинаудың таптырмас дереккөзі. Метабелгіленімдер қойылған корпустардан зерттеушілер өзіне қажетті стиль, кезең, автор т.б. ақпараттар бойынша мәлімметтерді 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.

Корпус мәтіндеріне метабелгіленімдер қою мәселесі А.Байтұрсынұлы атындағы Тіл білімі институтының Қолданбалы лингвистика бөлімінде 2015-2017 жылдардағы гранттық жоба бойынша зерттеліп келеді. Сонымен қатар 2016 жылы институт А.Фазылжановаңың жетекшілігімен «Қазақ тілінің ұлттық корпусын әзірлеу және жасау» атты мақсаттық жоба бойынша метабелгіленім әзірлемесін дайындады. Метабелгіленім енгізушілер үшін арнайы нұсқаулық жасалды және төмендегі тармақтарға негізделген әрбір стиль, корпус түрлеріне қойылатын

ұяшықтар құрастырылды. Ұяшық 23 тармақтан тұрады. Латын қаріпті корпус құрастыру тапсырмасы кезінде де осы метабелгіленім әзірлемесін программаға енгізу жұмысы жалғасын тапты. Сөйтіп, жартылай автоматты метабелгіленім енгізу бағдарламасы жасалды. Төменде осы бағдарламаның жұмыс үстелі ***1-суретте*** берілге



1-сурет. Мәтіндерге метабелгіленім енгізудің жартылай автоматты бағдарламасының жұмыс үстелі

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. Авторы анық емес мәтіндердің метабелгіленімдерінде автор деген ұяшық толтырылмай бос қалдырылады.

Автордың аты-жөнін беруде атын, әкесінің атын, фамилиясын толық жазу керек пе, әлде аты мен әкесінің атын қысқартып жазу керек пе дегенді де метабелгіленім жасауда алдымен шешіп алу қажет. Сонымен қатар оны фамилиядан бұрын беру немесе соңынан беру жағы да метабелгіленім қоюдағы бірізділік үшін қажет. Мысалы: Автор: А.Қайырбекова; Автор: Қайырбекова А.; Автор: Айжан Қайырбекова т.б.

Кейбір мәтіндерде авторға қатысты мәлімет берілмейді. Айталық, фольклорлық шығармалар ауызша тарағандықтан, белгілі бір авторы жоқ. Мұндайда автор туралы ақпарат берілетін кесте бос қалдырылады немесе авторы белгісіз деген сияқты белгіленім қоюға болады.

Авторға қатысты кейде қосымша тармақтар енгізуге тура келеді. Мәселен, ертегілер ауызша тараған мен, яғни авторы болмағанмен, ол ертегілерді жинақтап құрастырушы автор болады. Мұндайда метабелгіленімге «құрастырушы» деген ұяшық енгізуге болады. Сонымен қатар авторға қатысты ақпарат мынадай жағдайда өзгеріске түседі. Бұл корпусқа алынған мәтін аударма болған жағдайда кездеседі. Аударылған мәтінде түпнұсқадағы автормен қоса, оны аударған аудармашы аты туралы да ақпарат беру қажет.

Мәтін авторына қатысты тағы бір мәселе газет журналдарда берілген авторы жоқ хроникаларды беруде де туындайды. Мұндайда кейде авторы ретінде газет журнал редакторын алуға болады, яғни газет журналдағы мәтін авторы болмаған

жағдайда, газет редакторының аты және авторға қатысты мәліметтер (жынысы, жасы) көрсетіледі.

Метабелгіленімдер ұяшықтарында бір ғана авторға қатысты ақпаратты енгізуде әртүрлі стильдерге қатысты түрлі мәселе туындайды.

Ал кейбір мәтіндерде авторды алу немесе аудармашы, құрастырушыны алу мәселесін шешіп алу қажет. Кейбір аударма мәтіндерде авторды да, аудармашыны да көрсетуге болады.

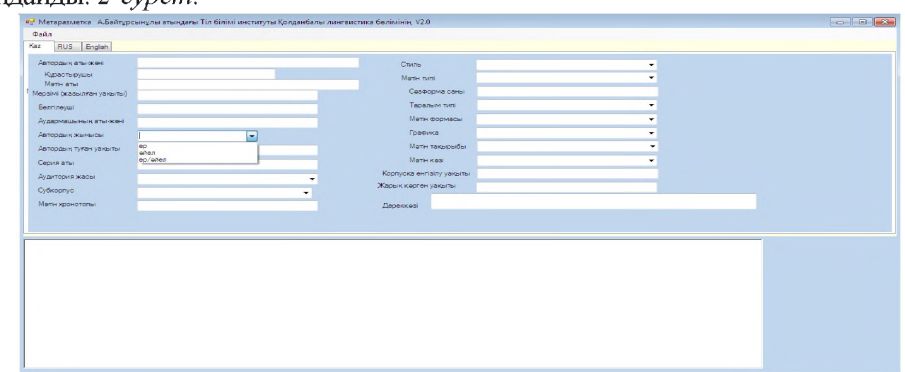
Метаразметканың авторға қатысты тағы бір түрі – автордың жас ерекшелігінің де көрсетілуі. Кейбір корпустарда автордың шығарманы жазған кездегі жас шамасы көрсетілсе (Британ, Чех), кейбір корпустарда автордың туған жылы, күні туралы нақты, дәл мәліметтер беріледі немесе шамамен көрсетіледі (Орыс тілінің ұлттық корпусы). Яғни автордың туған жылы, күні туралы нақты, дәл мәліметтер санмен беріледі. Ал автордың жасын анықтау қиын болған жағдайда «белгісіз» екендігі туралы белгі қойылады. Ал ұжымдық, жалпылама, белгісіз авторлар болған жағдайда жас ерекшелігі берілмейді немесе ұжымдық авторлар болған жағдайда «әртүрлі» деген белгіленім қоюға болады.

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. Автордың жас ерекшелігін көрсетуде оның жасын нақты көрсетпей, туған күні, айы, жылын жазуға да болады. Мысалы: 18.09.1973.

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. Автордың жынысы анық емес болған жағдайда, «белгісіз» деген белгіленім қойылады. Г.Қапан деген сияқты фамилия болғанда ер не әйел екенін белгілеу мүмкін емес, мұндайда белгісіз деген белгі қойылады не ұяшық бос қалдырылады.

«Жынысқа» қатысты ұяшықты толтыруда мәтін авторының, аудармашының, редактордың, құрастырушының қайсысының жынысын көрсетеміз деген мәселе туындайды.

Метабелгіленім бағдарламасында мұндай белгіленімдерді жазып отыру қиындыққа түспес үшін дайын форматтары тізімделеді. Белгілеуші солардың бірін таңдайды. *2-сурет.*



2-сурет. Жынысқа қатысты ұяшық

Башқұрт тілінің зерттеушісі З.А.Сиразитдинов автордың, информанттың ұлтында көрсеткен [1, 32 б.]. Ал орыс тілінің ұлттық корпусында ұлтқа қатысты белгіленім берілмейді.

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 немесе мақала атауы жоқ жай хроника ғана болған жағдайда «жоқ» деген белгіленім қойылады. Телеарна, радиодан алынған (ауызша не жазбаша) мәтін болса, телебағдарлама аты жазылады. Мәтін атауы дегенде басын ашып алатын мәселе кітап, жинақ мәтіндеріне қатысты да туындайды. Кітап, жинақ ішіндегі тақырыптар мәтін атауы ретінде беріле ме, әлде кітаптың, жинақтың аты мәтін атауы ретінде беріле ме? Метабелгіленімдер қоюда мәтін атауы ретінде кітап, жинақ ішіндегі тақырыпшалар алынады. Ал кітаптың, жинақтың сыртқы бетіндегі атауы метабелгіленімдер ішіндегі дереккөз (источник) деген ұяшықта беріледі. Бұл әртүрлі стиль мәтіндерінің бәріне ортақ ұстаным. Мәтін оқулық ішінен алынған тақырыптан алынса, тақырып атауы жазылады. Ал оқулық атауы дереккөзде көрсетіледі. Ғылыми мәтіндер мақала түрінде болса, мақала аты жазылады.

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.

Мерзімі (жазылған уақыты): 26.01.2016 жыл

Ал мәтіннің жазылу уақыты туралы нақты ақпарат болмаған жағдайда оның мерзімі 5-10 жылы аралағында шамамен алынады.

Мерзімі (жазылған уақыты): шамамен 1998-2000 жылдар

Кейде мәтіндердің жазылу уақыты туралы нақты мәлімет болмаған жағдайда корпусқа салынған мерзімі алынады. Мәтіннің корпусқа салынған мерзімі жазу кей жағдайда метабелгіленімдер қатарында арнайы жеке ұяшықта беріледі. Корпусқа енгізілу уақыты корпустың жасалу уақытымен бірдей болып келеді. Қазақ тілінің ұлттық корпусы қазіргі кездері ғана жасалып жатқандықтан, соңғы жылдар жазылады. Мысалы: 20.07.2016.

Мерзімге қатысты мәтіннің жазылған уақытынан басқа жарық көрген уақытын да метабелгіленімдердің арнайы ұяшығында беруге болады. Орыс тілі корпусында «Дата публикации» деген атаумен арнайы ұяшық берілген. Кейбір мәтіндер (шығарма, монография, оқулық т.б.) жазылған уақытында жарық көрсе, кейбірі кейін өңделіп қайта басылады. Бұл жерде кітаптың жарыққа шыққан, яғни қайта басылып шыққан уақыты көрсетіледі.

Мерзімі (жазылған уақыты): 1963 жыл

Жарық көрген уақыты: 1991 жыл

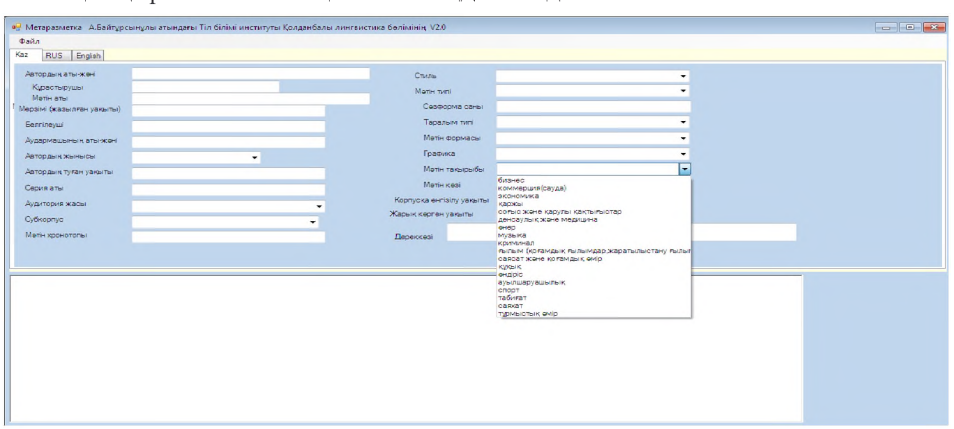
Метабелгіленімдердің бірі – корпусқа енгізілген мәтіндердің әрбіріндегі **сөзқолданыс саны** туралы ақпарат. Корпусқа мәтін енгізуде стильдер арасалмағы негізінен теңгерімді болғанмен, кейбір жанрларға қатысты мәтіндер көлемі әртүрлі

болып келеді. Мысалы хабарландырулар, құттықтаулар, жаңалықтар т.б. өте қысқа, олардағы сөзқолданыс саны да он шақты сөзден тұруы мүмкін. Сондықтан корпуска енгізілген метабелгіленімдердегі мәтіндердегі сөзқолданыс саны оннан он мыңдаған сөзқолданысқа жетеді. Сөзқолданыс саны өте үлкен болып келетіндері – көбінесе корпуска 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. Жеке файлда тұрған әрбір мәтінді ашқанда компьютер экранының астыңғы жағында сол мәтіндегі сөз саны шығып тұрады, осы санды метабелгіленімге енгізу керек. Сөзқолданыс санын мәтінді енгізу барысында арнайы компьютерлік бағдарламалар арқылы анықтайды. Кейбір корпустарда мәтіндердегі сөйлемдер саны туралы да ақпараттар берілген.

Корпустарда сөзқолданыс терминінің орнына сөзформа деп қолдану да кездеседі.

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*.

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. Төмендегі ***3-суретте*** тақырыптар әртүрлі. Белгілеуші корпуска енгізіліп отырған мәтіннің тақырыбын осының ішінен таңдап алады.

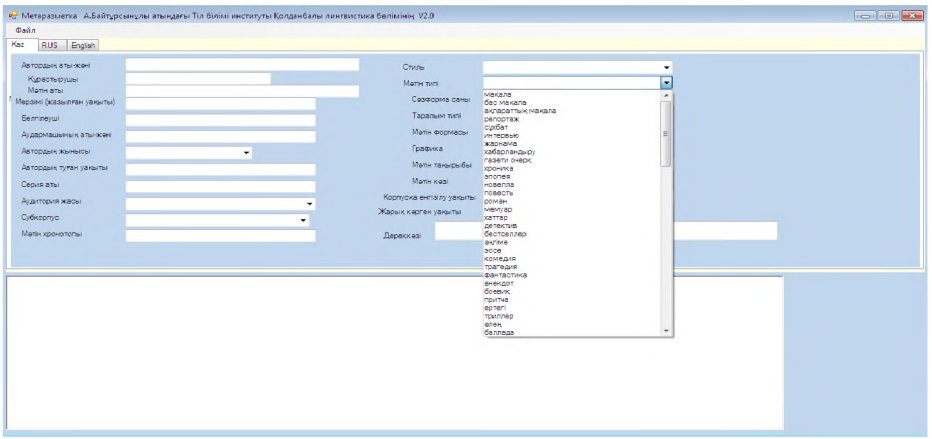


3-сурет. Мәтін тақырыбы ұяшығы

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 жж. немесе Қазақстан, Кеңес өкіметі жылдары т.б. Кейбір мәтіндерде мәтіндегі оқиғаның мерзімін, яғни хронотопын анықтау мүмкін болмайды. Оқиғағаға құрылмаған мәтіндер де болуы мүмкін. Мұндайда метабелгіленімдер тармақтарындағы хронотоп ұяшығына «бейтарап» деген белгі қоюға болады немесе мәтінде (көркем шығармада) суреттелген оқиғаның мерзімі мен орнын анықтау мүмкін болмаған жағдайда, «белгісіз» деген белгіленім қойылады.

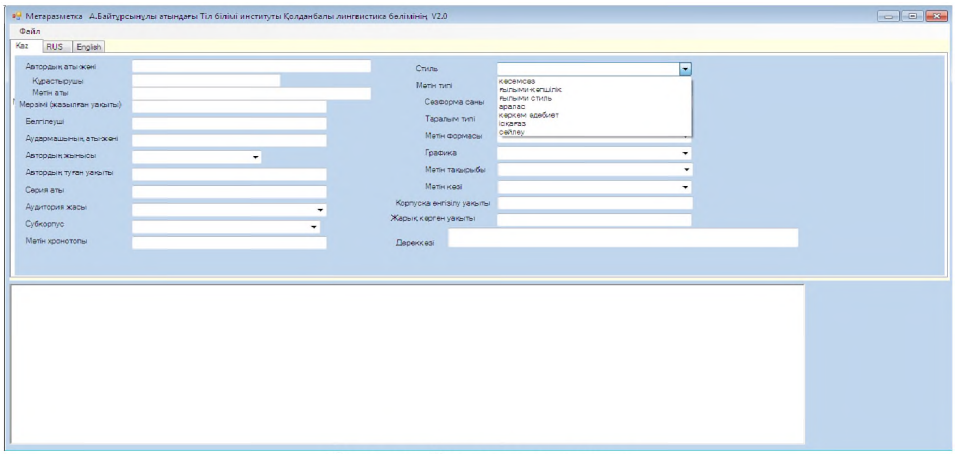
Келесі бір метабелгіленімдердің бір түрі – **мәтін типі.** Мұнда мәтіннің белгілі бір жанрға қатысы көрсетіледі. Мысалы: ғылыми стильдегі мәтін болса, мынадай типтері көрсетіледі; *мақала, монография, оқулық, реферат, оқу құралы, тезис, аңдатпа, түйіндеме, баяндама, пікір* т.б.Публицистикалық стиль бойынша публицистикалық мәтіннің белгілі бір жанрға (типке) қатысы көрсетіледі. *күнделік, репортаж, интервью,мақала, бас мақала, ақпараттық мақала, сұхбат, жарнама, хабарландыру, газети очерк, хроника* т.б.Ресми стиль мәтіндері болса, типтері көрсетіледі: *өтініш, арыз, сенімхат, қолхат, түсініктеме, мәлімдеме, анықтама, мінездеме, шақырухат, хабарландыру, жарнама, қатынасқағаз, акт* т.б.Ауызша жазылып алған мәтіндер немесе ауызша дискурстан алынып, қағазға түсірілген әртүрлі стильдегі мәтіндер бейтарап стиль ретінде танылады да, типтері былай беріледі: *жаңалықтар, сұхбат, әңгіме, комментарий, репортаж, реплика* т.б.Көркем әдебиет стилінде көркем мәтіннің белгілі бір типке қатысы көрсетіледі. Олар: *эпопея, новелла, повесть, роман* т.б.Көркем мәтін поэзия жанры болса, оның да типтері көрсетіледі. Олар: *өлең, баллада, поэма, арнау, жыр, айтыс, терме, ән, қара өлең* т.б. Төменде мәтін типінің ұяшығы берілген. ***4-сурет.*** Белгілеуші осы ұяшықтан енгізіліп отырған мәтіннің типіне қарай қажетті белгіленімді таңдайды.



4-сурет. Мәтін типі ұяшығы

Жанр деген термин әдебиет саласында да өзіндік мәнге ие болғандықтан, онымен шатастырмас үшін корпус құрастыруда кейде оны «мәтін типі» деген терминмен де қолданады. 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,

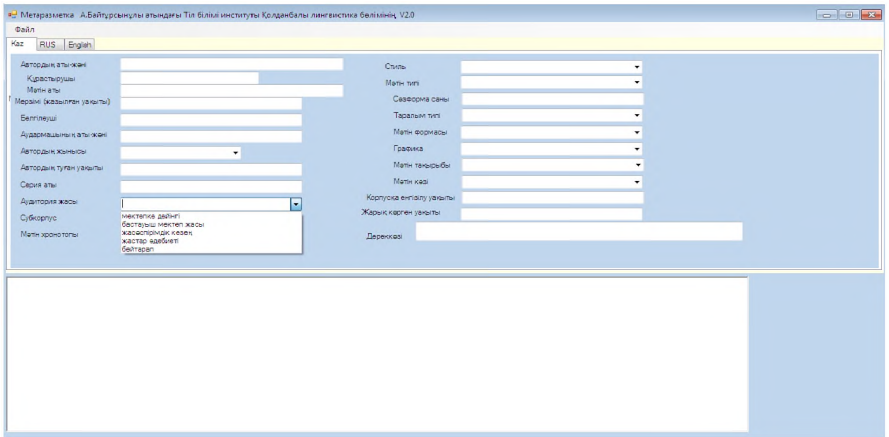
аймақтық, қарапайым, жеке-авторлық стильдерді беруге болады.Алайда мәтіндерге осылайша стилистикалық сипаттама жасау корпустарда көбінесе берілмейді. Оның орнына мәтін типі мен жанрына көбірек көңіл бөлінеді. Стильдерді метабелгіленім программасына енгізуде де белгілеуші осы төменде берілген ***5-суреттегі*** ұяшықтан стильдердің бірін таңдайды.



5-сурет. Стильдер ұяшығы

Метабелгіленімдердің келесі түрі – **аудиторияның жасын көрсету.** Мәтіннің кімге арналғанын білу мәтіннің мазмұны мен онда қолданылатын тілдік құралдар-ды да айқындайды. Мәтіндерге мұндай жас ерекшелігіне қарай белгіленім қою балалар әдебиетін, белгілі бір жас кезеңдеріне арналған оқулықтарды табуға мүм-кіндік 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. Әсіресе ересектерге ар-налған мәтіндерге белгіленім қойылмайды, аудиториясы бейтарап болып келеді. Мұндайда «бейтарап» деген белгіленім койылады. Аудитория жасына қатысты белгіленім коюда да параметрлерді алдын ала ойластырып алу қажет. Төменде метабелгіленім енгізудің жартылай автоматты бағдарламасының аудиторияның жасына қатысты параметрлер ұяшығы көрсетілген. ***6-сурет.***

Метабелгіленімдердің келесі түрі – мәтіннің кімге, бұл жерде **аудиторияның білім дәрежесіне** қарайғы ерекшелігін көрсету. Бұл жерде мәтін аудиториясының жалпы білімі немесе арнайы білімі, сондай-ақ жоғары білімі немесе біліміңің 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*. Алайда корпустардың көпшілігінде аудиторияның білім дәрежесі көрсетіле бермейді. Ол бір жағы аудиторияның білім дәрежесін анықтаудың кейде қиындық тудыруына байланысты.



6-сурет. Аудиторияның жасына қатысты ұяшығы

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 қолдап сканир жасалады немесе қолдан теріледі. Жарық көрмеген мәтіндер қолжазба ретінде көрсетіледі. Интернеттен алынған мәтіндерде сайт аты беріледі. Жарық көрген газет-журналдардың шыққан уақыты алынады.

Метабелгіленім параметрлерінің ішіндегі ең көлемді ақпарат берілетіні де дереккөз ұяшығы болып табылады. Онда кітап, жинақ аты, жылы, баспасы, жарық көрген жері т.б сияқты толық ақпарат беріледі. Мәселен, көркем шығарма мәтіні болса, дереккөзі ретінде ол туралы толық ақпарат (жинақ (кітап) аты, жері, баспа, жылы) жазылады.

Дереккөзі: Әуезов М. Абай жолы. 2 том. – Алматы, Ғылым, 1947.

Мәтін монография (ғылыми, іскери т.б.) ішінен алынған болса, монография туралы толық ақпарат (жері, баспа, жылы) беріледі.

Дереккөзі: Жаңабекова А. Функционалды грамматиканың метатілі. – Алматы, Дайк-пресс, 2012. – 150 бет.

Мәтіндер жарық көрген **баспа атын да көрсету** – метабелгіленімнің бір түрі. Бұл әсіресе кітаптарға (монография, оқулық, оқу құралы, шығарма т.б.) арналған. Баспа атына қатысты мәлімет кейбір корпустарда арнайы жеке ұяшықта берілмейді, дереккөздерінде басқа да ақпараттармен бірге көрсетіледі.

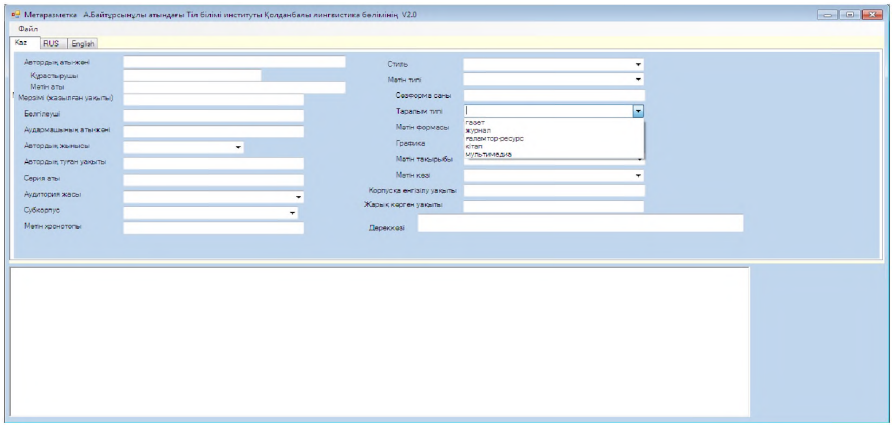
Корпусқа енгізілген мәтіннің корпусқа енгізуге дейін қандай формада (электронды, кітап, газет-журнал, іскери құжат т.б.) болғанын көрсету де қажет. Мұны **тара-**

лым типі деп атайды. Мәтін газет-журналдан алынған болса, таралымы типі газет не журнал болады.

Таралым типі: газет/журнал

Мәтін интернет-басылым, блок, сайттардан алынған болса, таралымы типі интернет-ресурс болады.

Таралым типі: ғаламтор-ресурс. Төменде *7-суретте* таралым типінің түрлері ұяшықта көрсетілген.



7-сурет. Таралым типінің ұяшығы

Таралым типіне ұқсас метабелгіленімнің бірі – **мәтін көзі.** Корпусқа жинақталатын мәтіндер әртүрлі жолмен жинақталады. Мәтін (көркем шығарма, өлең, газет, журнал т.б.) интернеттен алынған болса, дереккөзі ретінде «интернет» (ғаламтор) деген белгіленім қойылады.

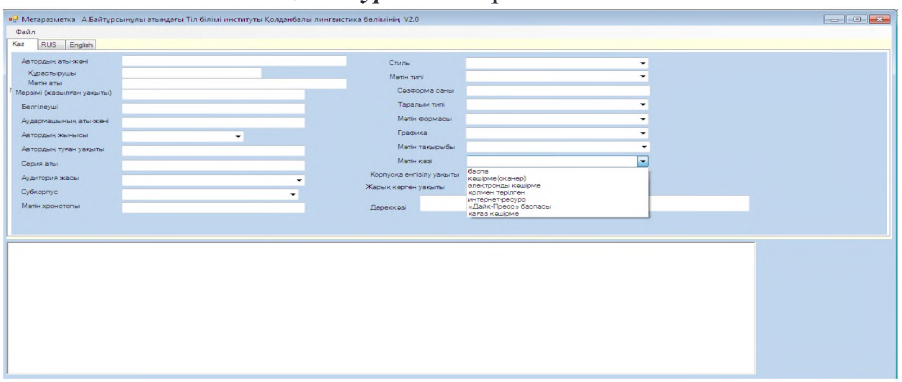
Мәтін көзі: ғаламтор-ресурс

Мәтін (көркем шығарма, өлең, газет, журнал т.б.) баспадан алынған болса, дереккөзі ретінде «баспа аты» көрсетіледі.

Мәтін көзі: «Дайк пресс» баспасы

Баспадан алынған, бірақ баспа атын анықтау қиындық тудырған жағдайда «баспа» деген белгіленім қойылады.

Мәтін көзі: баспа. Төменде ***8-суретте*** берілген.

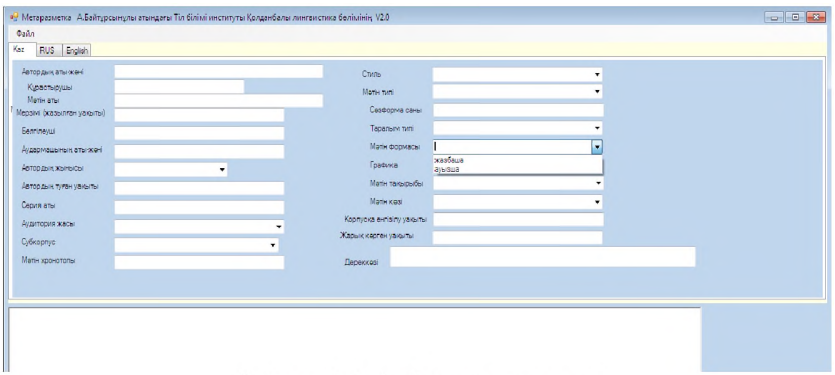


8-сурет. Мәтін көзі ұяшығы

Корпусқа енгізілетін мәтіндер әртүрлі пішінде (ауызша, жазбаша) болуы мүмкін. Мұны **мәтін формасы** деп атайды. Газет-журналдар, көркем шығарма, монография, оқулық т.б. мәтіндері жазбаша мәтіндер болғандықтан, «жазбаша» деген белгіленім қойылады.

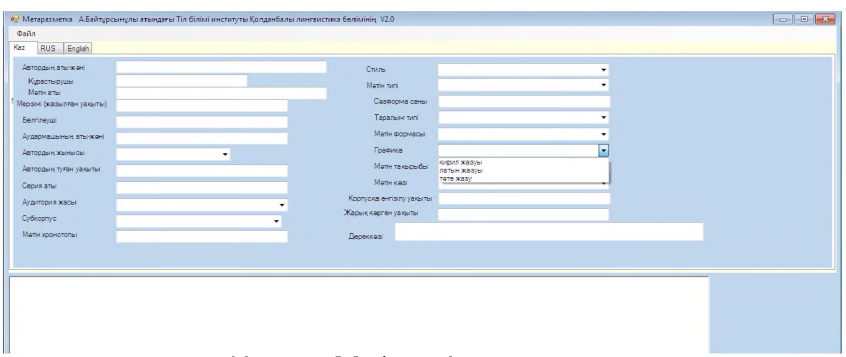
Мәтін формасы: жазбаша

Видео/аудиожазбалар болған жағдайда «ауызша» деген белгіленім қойылады. Мәтін формасы: ауызша. Мәтін формасы ұяшығы ***9-суретте***.



9-сурет. Мәтін формасы ұяшығы

Мәтінді сипаттауда тағы бір берілетін ақпарат оның қандай **қаріптермен** жазылғандығы туралы ақпарат. Қазақ тіліндегі мәтіндер төте, латын, кирил әріптерімен жазылғандықтан, алынған мәтіннің графикасын көрсетіп отыру да қажеттік туғызады. Алайда қазіргі мәтіндер негізінен кирил қарпімен жазылған. Корпусқа негізінен осы кирил қарпімен алынған мәтіндер алынады. Ал тарихи корпустар үшін төте, латын қаріптерімен жазылған мәтіндерді де алу қажет. Төменде ***10-суретте*** мәтіннің графикасына қатысты ұяшық көрсетілген.



10-сурет. Мәтін графикасы ұяшығы

Корпустарды іштей стиль бойынша, формасы бойынша т.б. критерийлер бойынша шағын корпустарға бөлуге болады. Осыған орай метабелгіленімдер ұяшығында алынған мәтіннің негізінен қандай корпусқа жататыны жайлы ақпарат беріледі. Мұндай шағын корпустарды орыс тілінде «подкорпус» терминімен атайды. Қазақ тілінде **субкорпус** ретінде көрсетуге болады. Олар: газет мәтіндері

субкорпусы, ауызша субкорпус, мультимедиялық субкорпус, көркем мәтіндер субкорпусы, поэтикалық мәтіндер субкорпусы, ресми мәтіндер субкорпусы, т.б. Орыс тілінде осы шағын корпустардың жиынтығын «негізгі корпус» деп атайды. Ал осы субкорпустардың барлық жиынтығын «Ұлттық корпус» деп атайды.

Публицистикалық стиль корпусына жазбаша және ауызша формадағы газет-журнал және радио/теледидар мәтіндері кіреді. Газет-журналдан алынған мәтін болса, газет мәтіндері корпусына жататындығы көрсетіледі.

Субкорпус: газет мәтіндері

Радио/теледидардан алынған аудиожазбалар болса, ауызша корпусқа жататындығы көрсетіледі.

Субкорпус: ауызша

Теледидардан алынған видеожазбалар болса, мультимедиялық корпусқа жататындығы көрсетіледі;

Субкорпус: мультимедиялық

Көркемсөз стилінде жазылған мәтіндер дегенге көбінесе проза/драма жанры енеді;

Субкорпус: көркем мәтіндер

Поэзия жанры «Корпус поэтических текстов» деген атаумен жеке субкорпус болады.

Субкорпус: поэтикалық мәтіндер

Оқулықтардан алынған мәтіндер болса, оқыту корпусы ретінде жазуға болады.

Субкорпус: оқыту

Ғылыми стильде жазылған мәтіндерді субкорпус ретінде шығаруға болады.

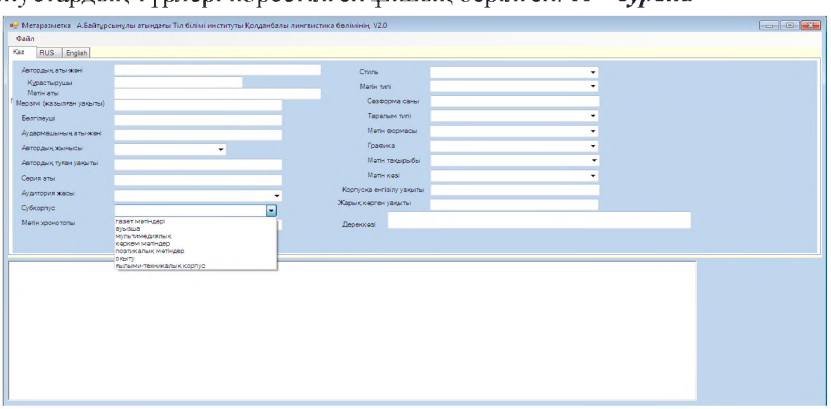
Субкорпус: ғылыми-техникалық корпус

Іскағаздар, ресми стиль мәтіндері «ресми мәтіндер субкорпусы» ретінде дербес корпус бола алады.

Субкорпус: ресми мәтіндер

Негізгі корпуста жоғарыда аталған субкорпустардың барлығы қамтылады. Негізгі корпусқа қатысты белгіленім жеке субкорпустар жасалып біткеннен кейін, яғни негізгі корпусқа біріктірілгеннен кейін қойылады.

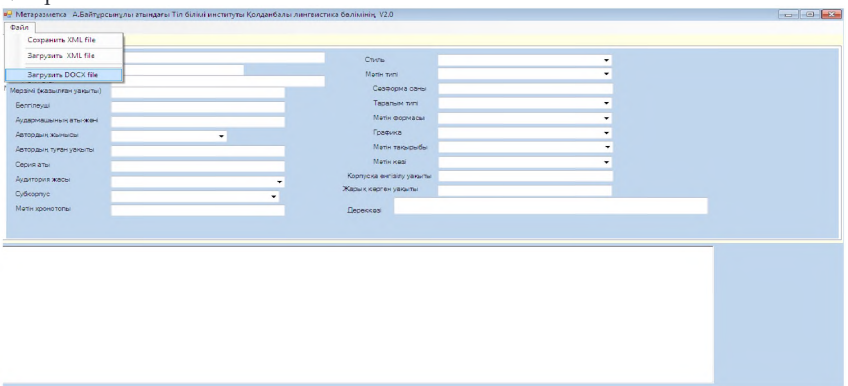
Субкорпус: негізгі корпус. Төменде метабелгіленім енгізу бағдарламасынан субкорпустардың түрлері көрсетілген ұяшық берілген. ***11 - сурет.***



11-сурет. Субкорпус ұяшығы

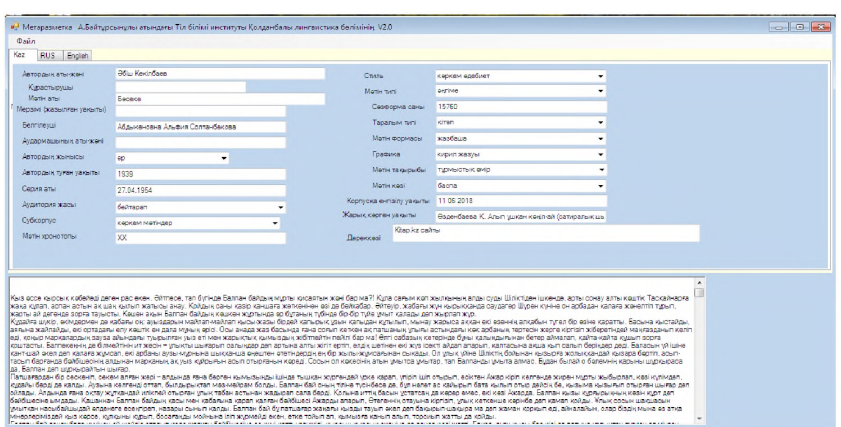
Орыс тілінің ұлттық корпусында метабелгіленім енгізуші адамның аты да ұяшықта (разметчик) берілген. Қазақ тілінің ұлттық корпусын жасауда да **белгілеуші** адам туралы ақпарат беруге болады.

Жоғарыда аталған метабелгіленім түрлері бойынша корпусқа енгізілетін мәтіндерге белгілеуші (разметчик) метабелгіленімдерді осы жартылай автоматты метабелгіленім қою бағдарламасы арқылы енгізеді. Бұл үшін алдыменэлектронды мәтіндердегі әр шығарманы жеке файлға сақтап, оларды авторы бойынша жеке папкаға жинаймыз, яғни Word файлдарды бір папкаға, ал XML файлдарды жеке папкаға сақтаймыз. Мәтіннің авторы, жазылған уақыты, алынған дереккөзі, стилі т.б. 23 параметр бойынша мәліметтер метабелгіленім бағдарламасына толтырылады. Сосын «Загрузить DOCX file»деген нұсқаулық арқылы электронды мәтінді жүктеп, «Сохранить XML file» нұсқаулығы арқылы метабелгіленімді сақтаймыз. ***12-суретте*** метабелгіленімдер мен мәтіндерді программаға (корпусқа) жүктеу жолы көрсетілген ұяшық берілген.



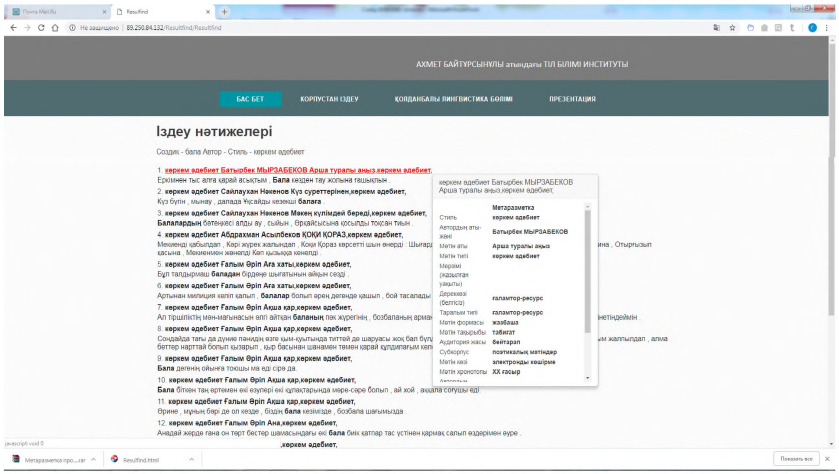
12-сурет. Метабелгіленімдер мен мәтіндерді программаға (корпусқа) жүктеу жолы

Төмендегі ***13-суретте*** 23 параметр бойынша метабелгіленім толық енгізілген және сол мәтін бірге тіркеліп сақталған ұяшық берілген. Осы процесс орындалғаннан кейін метабелгіленім қойылған мәтіндер корпусқа еніп отырады.



13-сурет. Метабелгіленім толық енгізілген және сол мәтін бірге тіркеліп сақталған ұяшық

Қазіргі кезде осы 23 параметрмен Қолданбалы лингвистика бөлімінде мәтіндер корпусы (14-сурет) жұмыс жасайды.



14-сурет. А.Байтұрсынұлы атындағы Тіл білімі институты Қолданбалы лингвистика бөлімінің мәтіндер корпусындағы метабелгіленім терезесі

Сонымен, корпус жасауда метабелгіленімдерді енгізу қажеттігі дау тудырмайды. Олар корпусты пайдаланушыға, әсіресе тілші-зерттеушілерге кез келген қажетті ақпаратын тез әрі оңай тауып алуға мүмкіндік береді.

ӘДЕБИЕТТЕР ТІЗІМІ:

[1] Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш., Ибрагимова А.Д. Информационные системы и базы данных башкирского языка. – Уфа: Книжная палата РБ, 2013. – 116 с.