Хәкім Абай халқымыздың ұлттық дүниетанымындағы күнкөріс психологиясын еңбек пен ғылымды игеру арқылы ғана мал табуға болады дей отырып назарға алады. Абай келешек ұрпақтың терең ойлы, қайырымды, иманды, инабатты ұл мен қыз болуын, олар ұстанатын қағиданы айтып кеткен. Абай "Менің сөзімді біреу ұғар, біреу ұқпас",-дей келе, осыны ұғатын ұрпаққа тіл ұстартып, өнер шашып кеткен ақын.

Қорытынды.Ақынның мәңгі өшпес мұраларының өміршеңдігін мына пікірмен түйін жасағымыз келеді. "Абайдың шығармалары мен көркем образдары онда суреттелген дәуірмен бірге өшіп қалған жоқ. Олардың әсер етушілік қуаты – өз тұсындағы халық өмірін Абайдың мейлінше кемелденген көркем формада бейнелеуінде" [10.,18]. Әр заманның өзіндік саясаты болар, қандай жағдайда да ұлтымызды парасат биігінде қалуға, адалдыққа, абзал кісілікке шақырған.

Ғибратты туындылары арқылы келер ұрпаққа өнегелі ой, ғибратты үлгі, өмір ережесін, тәрбие тағылымын ұсынған.

Пайдаланылған әдебиеттер тізімі:

1 Қасым-Жомарт Тоқаев. Абай және XXI ғ. Қазақстан. Қазақ әдебиеті газеті, № 1-2 саны, 10 қаңтар, 2020 жыл, 5 б.-газет

2 Торайғыров С. Таңдамалы шығармалар жинағы. – Алматы: Жазушы,1989.–321 б.-кітап

3 Баласағұн Ж. Құтты білік. – Алматы: Жазушы, 1986. – 616 б.-кітап

4 Абай (Ибрагим Құнанбаев) Екі томдық шығармалар жинағы.– Алматы: Жазушы, 1986. – 304 б.-кітап

5 Мырзахметов М. Абайтану тарихы. – Алматы: Ана тілі, 1994. – 192 б.-кітап

6 Новай А. X том. – Ташкент: Баспа, 1970. – 450 б.-кітап

7 Тебегенов Т. Абайтану ұлағаты. – Алматы: Ұлағат, 2013. – 204 б.-кітап

8 Психология-Адамзат ақыл-ойының қазынасы. 10 томдық,7 том–Алматы: Таймас, 2005. – 480 б.-кітап

9 Абай қара сөздері. – Алматы: Ел , 1992. – 272 б-кітап

10 Абай энциклопедия. – Алматы:Қазақ энциклопедиясы, Атамұра, 1995.-720 б.-кітап.

References:

1. Qasym-Jomart Toqaev. Abai jäne HHI ğ. Qazaqstan. Qazaq ädebietı gazetı, № 1-2 sany, 10 qañtar, 2020 jyl, 5 b.-gazet
2. Toraiğyrov S. Tañdamaly şyğarmalar jinağy. – Almaty: Jazuşy,1989.–321 b.-kıtap
3. Balasağūn J. Qūtty bılık. – Almaty: Jazuşy, 1986. – 616 b.-kıtap
4. Abai (İbragim Qūnanbaev) Ekı tomdyq şyğarmalar jinağy.– Almaty: Jazuşy, 1986. – 304 b.-kıtap
5. Myrzahmetov M. Abaitanu tarihy. – Almaty: Ana tılı, 1994. – 192 b.-kıtap
6. Novai A. H tom. – Taşkent: Baspa, 1970. – 450 b.-kıtap 7
7. Tebegenov T. Abaitanu ūlağaty. – Almaty: Ūlağat, 2013. – 204 b.-kıtap
8. Psihologia-Adamzat aqyl-oiynyñ qazynasy. 10 tomdyq,7 tom–Almaty: Taimas, 2005. – 480 b.-kıtap
9. Abai qara sözderı. – Almaty: El , 1992. – 272 b-kıtap
10. Abai ensiklopedia. – Almaty:Qazaq ensiklopediasy, Atamūra, 1995.-720 b.-kıtap.

МРНТИ 11.25.41 https://doi.org/10.51889/2020-4.1728-7804.60

Пирманова К.К.,[1] Карбозова Б.Д.,[2] Токмырзаев Д.О.[3]

[1]Казахский национальный университета имени аль-Фараби,
Алматы, Казахстан
[2,3]Институт Языкознания имени А.Байтурсынова,
Алматы, Казахстан

ТЕХНОЛОГИЯ ПРОГРАММЫ ПОЛУАВТОМАТИЧЕСКОЙ МЕТАРАЗМЕТКИ И ПРИНЦИПЫ АВТОМАТИЗАЦИИ МОРФОЛОГИЧЕСКОЙ РАЗМЕТКИ ТЕКСТОВ НАЦИОНАЛЬНОГО КОРПУСА КАЗАХСКОГО ЯЗЫКА

Аннотация

При изучении казахского языка необходимо уделять большое внимание области корпусной лингвистики и изучать ее теоретические и практические аспекты мирового уровня. В специальных изданиях научных журналов также публикуются статьи по общим и конкретным вопросам, касающимся создания и работы текстовых корпусов по всему миру. Однако известно, для казахского языкознания требуется специальное изучение многих вопросов, связанных с корпусной лингвистикой. Оно включает в себя: определение корпусной лингвистики и ее основных

понятий, место корпусной лингвистики в структуре языкознании, методы и т.д. Теоретические и практические аспекты вышеупомянутой корпусной лингвистики также должны быть приняты во внимание при создании базы данных текстов на казахском языке на основе компьютерного корпуса. Если корпусная лингвистика формируется как особый раздел казахского языкознания, она позволит многим специалистам по казахскому языку использовать крупномасштабные экспериментальные материалы, находить необходимые языковые данные и вносить соответствующие изменения. Все это способствует новому взгляду на эмпирические подходы к достоверности исследований казахского языка и внедрению наиболее важных языковых материалов в области науки.

Ключевые слова: корпус, национальный корпус казахского языка, лингвистические разметки, мета разметка, морфологическая разметка

Pirmanova K.,[1] Karbozova B.,[2] Tokmyrzhayev D.[3]

[1]al-Farabi Kazakh national university
Kazakhstan, Almaty,
[2,3] A.Baitursynov Institute of Linguistics,
Kazakhstan, Almaty,

THE TECHNOLOGY OF SEMI-AUTOMATIC META-MARKUP PROGRAM AND THE PRINCIPLES OF AUTOMATING THE MORPHOLOGICAL MARKUP OF THE KAZAKH LANGUAGE NATIONAL CORPUS TEXTS

Abstract

When studying the Kazakh language, it is necessary to pay great attention to the field of corpus linguistics and study its theoretical and practical aspects of the world level. Special editions of scientific journals also publish articles on General and specific issues related to the creation and operation of text corpora around the world. However, it is known that Kazakh linguistics requires special study of many issues related to corpus linguistics. It includes: the definition of corpus linguistics and its main concepts, the place of corpus linguistics in the structure of linguistics, methods, etc. Theoretical and practical aspects of the above-mentioned corpus linguistics should also be taken into account when creating a database of texts in the Kazakh language based on a computer corpus. If corpus linguistics is formed as a special section of Kazakh linguistics, it will allow many specialists in the Kazakh language to use large-scale experimental materials, find the necessary language data and make appropriate changes. All this contributes to a new look at the empirical approaches to the reliability of research in the Kazakh language and the introduction of the most important language materials in the field of science.

Keywords: corpus, national corpus of the Kazakh language, linguistic markings, meta-marking, the morphological marking

Пирманова К.К.,[1] Карбозова Б.Д.,[2] Токмырзаев Д.О.[3]

[1]Әл-Фараби атындағы Қазақ Ұлттық университеті,
Алматы, Қазақстан
[2,3]А.Байтұрсынұлы атындағы Тіл білімі институты,
Алматы, Қазақстан

ҚАЗАҚ ТІЛІ ҰЛТТЫҚ КОРПУСЫ МӘТІНДЕРІНІҢ ЖАРТЫЛАЙ АВТОМАТТЫ МЕТАБЕЛГІЛЕНІМ БАҒДАРЛАМАСЫНЫҢ ТЕХНОЛОГИЯСЫ МЕН МОРФОЛОГИЯЛЫҚ БЕЛГІЛЕНІМ ПРИНЦИПТЕРІ

Аңдатпа

Қазақ тілін оқу барысында корпустық лингвистика саласына көп көңіл бөліп, оның әлемдік деңгейдегі теориялық және практикалық аспектілерін зерттеу қажет. Ғылыми журналдардың арнайы басылымдарында сондай-ақ бүкіл әлем бойынша мәтін корпустарын құруға және олардың жұмысына қатысты жалпы және нақты мәселелер бойынша мақалалар жарияланады. Алайда, қазақ тіл білімі үшін корпустық лингвистикамен байланысты көптеген мәселелерді арнайы зерттеу қажет. Ол: корпустық лингвистиканы және оның негізгі ұғымдарын анықтауды, тіл білімі құрылымындағы корпустық лингвистиканың орнын, әдістер мен тәсілдерді және т. б. қамтиды. Жоғарыда аталған корпустық лингвистиканың теориялық және практикалық аспектілері компьютерлік корпус негізінде қазақ тіліндегі мәтіндердің деректер базасын жасау кезінде назарға алынуы тиіс. Егер корпустық лингвистика қазақ тіл білімінің ерекше бөлімі ретінде қалыптасса, ол қазақ тілін пайдаланушы әрі қазақ тілінде зерттеулермен айналысатын көптеген мамандарға үлкен көлемді тәжірибелік материалдарды пайдалануға, қажетті тілдік деректерді табуға және тиісті өзгерістер енгізуге мүмкіндік береді. Осының барлығы қазақ тілін зерттеудің шынайылығына эмпирикалық көзқарасқа және ғылым саласындағы ең маңызды тілдік материалдарды енгізуге ықпал етеді.

Түйін сөздер: корпус, қазақ тілінің ұлттық корпусы, лингвистикалық белгіленімдер, мета белгіленім, морфологиялық белгіленім

Введение. Корпуса делятся на различные подкорпусы в зависимости от их функции. Создание аннотированных корпусов, особенно лингвистических разметок для лингвистических исследований, очень важно и полезно. В зависимости от того, поставлена эта разметка или нет, корпуса делятся на аннотированные и не аннотированные. Даже если разметки корпуса важны для лингвистических исследований, прежде всего, необходимо отсортировать тексты в приложении и предоставить точную информацию о них.

Типы разметок, которые прикреплены к корпусу, можно разделить на две категории: лингвистические и внешние. Вне лингвистические разметки включает в себя:

1) разметка, которая описывает функции форматирования текста (темы, абзацы, пробелы и т. д.);

2) разметка, которая описывает автора и текст.

При этом информацией об авторе может быть не только его имя, но и его возраст, пол, годы его жизни и т. д. Текстовая информация, как правило, отличается от темы, на каком языке она написана, год, место издания, название и т. д. Наличие такой информации в корпусе позволяет искать текстовую поисковую систему и, в то же время, создавать необходимый инструмент для идентификации соответствующего документа. Иногда это называется экстралингвистическими разметками, а также используется как метатекстовая разметка или метаразметка. Есть также те, кто определяет лингвистическую разметку как внешний тег, а метаразметку внутренним тегом. Они включают текст и информацию об авторе: автор, название, год, место происхождения, жанр текста, тему, стиль, размер и т. д. Их также делят на библиографические, типологические, тематические, социальные, формальные (тексты, главы, части, абзацы, предложения и т. д.) и технические (исполнители, источники, извлеченные из электронных версий, даты обработки, закодированное время и т. д.).

Методика. У любого текста должен быть автор. Это:

а) если у текста есть автор, будет указано его полное имя; б) В случае нескольких авторов даются имена коллективных авторов. Такие как коллективные монографии, статьи в соавторстве и т.д .; в) обобщенный автор, такие тексты (т. е. документы, письма, тексты), написанные не отдельным лицом, а коллективом, учреждением; г) некоторые авторы текста могут быть неизвестны. Такие случаи особенно встречаются в газетах и журналах. Авторов таких текстов иногда отмечают условными именами. В случае если у текста нет автора, или же если он неточный, то в метаразметке ячейка предназначенная для автора не заполняется.

При разработке метаразметки нужно решить вопрос: будут ли писать только фамилию автора, необходимо ли написать полное имя, отчество, фамилию или имя и отчество, чтобы записать его в метаразметке. Также необходимо дать ему фамилию до или после, это нужно для последовательности метаразметки. Например: А.Кегенбекова; Автор: Кегенбекова А.; Автор: Алтын Кегенбекова и т.д.

Некоторые тексты не предоставляют информацию об авторе. Например, фольклорные сочинения распространяются устно, а автора нет. В этом случае таблица, в которой представлена информация об авторе, может быть оставлена пустой или вы можете пометить ее как неизвестного автора. Иногда приходится добавлять дополнительные пункты в отношении автора. Например, хотя сказки распространяются устно, есть автор, который собирает все эти сказки. В этом случае вы можете ввести пункт под названием “составитель” в метаразметку.

Кроме того, информация, относящаяся к автору, может быть изменена в следующих случаях. Если данный текст в корпусе переведен. В переведенном тексте также необходимо предоставить информацию об имени переводчика, в том числе об оригинальном авторе.

Еще одна проблема, связанная с автором текста, заключается в том, что при хронологии во многих газетах и журналах нет никаких авторов. В этом случае автор может иногда быть редактором газетного журнала, то есть имя редактора газеты и есть информация об авторе (пол, возраст).

При записи информации об одном авторе в ячейки метаразметок возникают различные проблемы с разными стилями. А в некоторых текстах необходимо найти автора или решить вопрос с переводчиком и состовителем. Некоторые переводческие тексты могут включать как автора, так и переводчика.

Другой тип метаразметки, связанных с автором – это возраст автора. В некоторых корпусах указывается возраст автора при написании произведения (Британский, Чешский), в некоторых корпусах приводятся точные сведения о дате рождения автора или указываются приблизительно (Национальный корпус русского языка). То есть точные сведения о дате рождения автора приводятся цифрами. А в случае, когда трудно определить возраст автора, делается разметка о том, что “неизвестно”. При наличии коллективных, обобщенных, неизвестных авторов возраст не присваивается или при наличии коллективных авторов можно поставить разметку “разное”.

Обсуждение. Иногда авторы отдельных характерных текстов, таких как дневник, личные письма, ставят имена, не давая их фамилий, даются под условным названием, но с указанием пола и возраста. В показе возрастных особенностей автора можно записать число, месяц, год рождения без четкого указания его возраста. Например: 14.06.1973.

В метаразметках иногда дается информация, касающаяся пола авторов. Возможно, что автор является женщиной или является мужчиной или не может быть четко выраженным полом. Обычно пол автора указывается в том случае, если автор текста является одним, а в коллективных текстах не указывается пол автора. В случае неясности пола автора ставится разметка “неизвестно”. При наличии такой фамилии, как А.Омар, невозможно установить мужчину или женщину, при этом делается разметка о неизвестности либо ячейка остается пустой. При

заполнении ячейки, относящейся “к полу”, возникает вопрос, какой из авторов текста, переводчика, редактора, составителя мы указываем пол.

В программе метаразметки описываются готовые форматы, чтобы не попасть в трудность записи таких разметок. Разметчик выбирает один из них.

Исследователь башкирского языка З.А. Сиразитдинов указал нацию автора и информанта [1, с. 32]. А в Национальном корпусе русского языка обозначение в отношении нации не выдается.

Название текста, внесенного в корпус, также является одним из основных метаразметок. У текста, введенного в корпус, могут быть не все заголовки. Если в тексте дано название тем, они закладываются в систему метаразметок, а названия текстов, под которыми не даны заголовки, не указываются. Это, как правило, короткие тексты в газетах и журналах, передаваемые внутри одной рубрики, следовательно, все тексты, помещенные в корпус, описывают использование естественного языка, даже если заголовок отсутствует, но в системе метаразметок не указывается или ставится отметка “нет” при наличии только простой хроники, не имеющей названия статьи. При наличии на телевидении, радио (устного или письменного) текста, записывается название телепрограммы. Проблема, в которой название текста может раскрыть начало, возникает и в отношении текстов книг, сборников. В книге, сборнике дается ли название текста, или название книги, сборника присваивается как название текста? При проставлении метаразметок в качестве наименования текста изымаются подтексты внутри книги, сборника. А на оборотной стороне книги, сборника дается название в ячейке метаразметки источник. Это общая позиция для всех текстов различных стилей. Если текст получен из темы, полученной из учебника, то название заголовка записывается. А название учебника указывается в источнике. При наличии научных текстов в виде статьи, название статьи записывается.

Одна из метаразметок о тексте – время написания текста. Как правило, такие разметки извлекаются из сведений, оставленных автором в конце текста при написании произведения. Чаще всего время написания текста определяется библиографическими, биографическими исследованиями.

Дата (время записи): 21.01.2016 года

При отсутствии точной информации о времени подписки, ее срок будет составлять около 5-10 лет.

Срок (время написания): приблизительно 1998-2000 годы.

Иногда при отсутствии точных сведений о времени подписки текстов снимается срок постановки в корпус. Срок записи текста в корпус, в некоторых случаях, в ряду метаболизмов, передается в специальной отдельной ячейке. Время ввода в корпус совпадает со временем изготовления корпуса. В связи с тем, что Национальный корпус казахского языка формируется только в настоящее время, записываются последние годы. Например: 29.12.2016.

Кроме времени написания текста, относящегося к сроку, можно также указать время публикации в специальной ячейке метаразметок. В корпусе русского языка представлена специальная ячейка под названием “Дата публикации”. Некоторые тексты (сочинение, монография, учебник и т. д.) если в записанное время они будут опубликованы, то некоторые будут обработаны и перепечатаны. Здесь указывается время, когда книга вышла в свет, то есть переиздана.

Дата (время написания): 1963 год

Дата публикации: 1991 год

Одна из метаразметок – информация о количестве словосочетаний в каждом из текстов, внесенных в корпус. Хотя соотношение стилей при вводе текста в корпус в основном сбалансировалось, размеры текстов относительно некоторых жанров различны. Например объявления, поздравления, новости и т. д. очень короткое, количество слов в них также может состоять из десятков слов. Поэтому число словосочетаний в текстах, введенных в корпус, достигает десятков тысяч словосочетаний. Количество словосочетаний очень великое – объемные труды, чаще всего такие как романы или монографии, введенные в корпус в целом. Но иногда не все такие произведения и монографии могут быть получены, а только определенные разделы. Количество словосочетаний в текстах указывается в присвоенной им системе метаразметок. При открытии каждого текста, находящегося в отдельном файле, в нижней части экрана компьютера появляется цифра в тексте, которое необходимо ввести в метаразметку. Число словосочетаний определяется при вводе текста посредством специальных компьютерных программ. В некоторых корпусах также представлена информация о количестве предложений в текстах.

В корпусах вместо словосочетания встречается применение термина словоформа.

Теперь одним из видов метаразметок является область применения текста, которая является самой общей типологической характеристикой текста. Так, в Национальном корпусе русского языка представлено 8 функциональных сфер применения. Это: учебно-научные, производственно-технические, официально-деловые, публицистические, рекламные, религиозные, литературные, бытовые.

Разделение текстов на определенные тематические группы также относится к одной из метатекстовых разметок. Например, общественные науки, физика, биология, путешествия, спорт, природа, искусство, политика и т. д. Однако такое деление на тематические группы иногда обусловлено. Так как некоторые тексты рассматриваются в рамках нескольких тем, в отношении нескольких областей. Поэтому, когда определенный текст относится к тематической группе, они разделяются не только на одну сферу, но и на несколько областей. В мировых корпусах (корпус Браун, Национальный корпус русского языка и др.) в художественной литературе зачастую не показываются тематические группы.

Таким образом, типы метаразметок пополняются такими ячейками, как тип текста (здесь указывается отношение текста к определенному жанру); возраст аудитории (знание, кому адресован текст определяет содержание текста и используемые в нем языковые средства); специфика текста (в которой речь идет о степени знаний аудитории); количество (объем среды, использующей тексты); источник (это источники, в которых получен текст); стиль текста; хронотоп (поскольку выделение некоторых текстов по тематическим группам затруднено, в некоторых корпусах необходимо было указать время и место написания текста, то есть указывается место написания текста и отношение к определенному периоду) и т.д.

Еще одна представляемая в описании текста информация о том, какими шрифтами она написана. В связи с тем, что тексты на казахском языке написаны алфавитом, который основал А.Байтурсынов, латинскими, кириллическими буквами, необходимо указывать график полученных текстов.

Корпуса по внутреннему стилю, по форме и т. д. по критериям можно разделить на небольшие корпуса. Таким образом, дается информация о том, к какому корпусу относится в основном текст, полученный в ячейке метаразметки. Такие небольшие корпуса называются термином “подкорпус” на русском языке. На казахском языке можно показать как субкорпус. Это: субкорпус газетных текстов, устный субкорпус, мультимедийный субкорпус, субкорпус художественных текстов, субкорпус поэтических текстов, субкорпус официальных текстов и др. В русском языке комплект этих небольших корпусов называется “основной корпус”. А весь набор этих субкорпусов называют “Национальным корпусом”.

В Национальном корпусе русского языка имя разметчика также представлена в ячейке (разметчик). В создании Национального корпуса казахского языка можно также проинформировать о обозначенном человеке.

Морфология – наука, изучающая строение слов. Для морфологии не важно выражение лексического значения частиц слова, так как оно рассматривает все языковые единицы по формальной структуре. С точки зрения межличностного смысла содержание приобретает только форму грамматических значений. Например, предметные, критические, количественные значения слов также относятся к грамматическому значению. Ученый-корреспондент С. Исаев назвал эти значения общим грамматическим значением [2, с. 159]. При постановке морфологической разметки важно точно определить классы слов, сгруппированные через эти общие грамматические значения. Потому что полный морфологический образ слов непосредственно относится к общим грамматическим значениям слов, то есть к существительным. Если котировка по отношению к сторонам слов будет неправильно поставлена, то будет неправильно сформулирована формулировка следующих грамматических форм.

Морфологическая разметка в русском языке осуществляется в три этапа: анализ-анализ (парсинг); сортировка; снятие омонимии [3, с. 111]. Метод лингвистической (морфологической) разметки, создаваемый на первом этапе, осуществляется по программе MYSTEM. Эта программа основана на сведениях А.А. Зализняка в “Грамматическом словаре русского языка”, и здесь возможные варианты морфологических признаков каждого словаря записываются в скобках [4, с. 69].

Для автоматизации процесса разграничения (разделения) морфемной структуры казахских слов и словоформ необходимо учитывать их словообразовательные, словоизменяющие или же формообразующие функции. Выяснение двуфункциональности аффиксов позволит отличить их как самостоятельное слово и позволит включить их в базу основ слов (лексем) или, в другом случае, как грамматические формы слов – в базу порождаемых словоформ [5], [6], [7], [8], [9]. Далее, отделив последовательность словоизменительных аффиксов от словоформы, можно включить их в базу (список) размеченных словоизменительных окончаний.

Как известно, в аглютинативных языках присоединение к основе слов последовательности словоизменительных аффиксов имеет четкую форму. Это означает, что они могут быть присоединены не только к одной основе, а ко всем подобным основам слов, относящихся к одной части речи, образующей систему словоизменений с одинаковой структурой. Можно сказать, что такая система словоизменений функционирует для всех тюркских языков. Особенно заметно это проявляется для часто употребляемых слов, относящихся к существительным и глаголам. Поэтому словарь словоформ, составленный на основе словоизмененных аффиксов, в основном, состоит из словоформ, относящихся к указанным частям речи.

Результаты. По вышеперечисленным видам метаразметки к текстам, вводимым в корпус, разметчики вводят метаразетку через эту полуавтоматическую программу установки метаразметок. Для этого сначала сохраним каждое произведение в электронных текстах в отдельный файл и собираем их в отдельную папку по автору, то есть храним файлы Word в одну папку, а XML-файлы в отдельную папку. Автор текста, дата написания, источник получения, стиль и т. д. Данные по 23 параметрам заполняются в программу метаразметок. Затем загрузим электронный текст с инструкцией “Загрузить DOCX file” и сохраним метаразметку с помощью руководства “Сохранить XML file”.

Таким образом, принцип автоматизации морфологической разметки Национального корпуса казахского языка осуществляется в следующем порядке:

1-блок. В корпусную базу в качестве самостоятельной базы вводится словарь основ казахских слов в алфавитном порядке с заранее размеченной по составу (односложное, двухсложное или многосложное) и по частям речи.

Данная словарная база основывается на материалах 15-ти томного и сокращенного 1-томного толкового словаря казахского литературного языка, который охватывает 106 тыс. реестровых слов (лексем) с разметкой по частям речи. Этот отдельно взятый список из 106 тыс. реестровых слов был размечен нами еще и по составу (в

перспективе намечается включить туда еще и семантическую разметку). Для примера рассмотрим отрывок таких данных в виде ***Таблицы 1***:

Таблица 1

Реестровое слово	**Части речи**	**По составу**	**По способу образования**	**По методу образования**
Аба	зт./сущ.	простое	Основное	
абай-қоқай	зт./сущ.	сложное	производное	Аналитический
Абайтану	зт./сущ.	сложное	производное	Аналитический
абайтанушы	зт./сущ.	сложное	производное	Аналитический
абайшылдық	зт./сущ.	простое	производное	Синтетический
Абайым	зт./сущ.	простое	Основное	Лексика-семантический

2-блок. В корпусную базу в качестве самостоятельной базы вводится: список всевозможных словоизменяющих аффиксов, составленных для каждой части речи казахского языка в отдельности. При этом структура (или последовательность) словоизменяющих аффиксов в базе является заранее размеченной на грамматическом уровне. Для примера рассмотрим отрывок таких данных "Аффиксы образования словоформ от корневых и производных основ существительных с конечным мягким слогом" как показано в ***Таблице 2***.

Таблица состоит из 4-х столбцов. Во 2-ом столбце даны словоизменяющие аффиксы без разделения структуры, в 3-ем через знак "+" даются структуры, а в 4-ом столбце – к каждой части словоизменяющих аффиксов приписаны условные обозначения разметок в сокращенном варианте, т.е. показаны грамматические разметки структуры словоизменяющих аффиксов через сокращения.

Необходимо отметить, что такие списки словоизменяющих аффиксов в виде таблицы составлены для основ слов с различными конечными буквами и для всех частей речи казахского языка в отдельности.

Таблица 2.

№ п/п	**Словоизм. аффиксы для основ слов с конечными гласными (кроме *у, и)***	**Структура словоизменяющих аффиксов**	**Разметка состава словоизменяющих аффиксов**
	Примеры основ слов с конечным мягким слогом (кроме *у, и)*: *Әке, келі, мәсі, ...*		
1	**2**	**3**	**4**
1	*Дей*	*дей*	*дей/СФ*
2	*сіз*	*сіз*	*сіз/ЖФ*
3	*сіздің*	*сіз+дің*	*сіз/ЖФ+дің/ІС*
4	*сізге*	*сіз+ге*	*сіз/ЖФ+ге/БС*
--	---------	----------	----------
10	*сіздер*	*сіз+дер*	*сіз/ЖФ+дер/КЖ*
11	*сіздерде*	*сіз+дер+де*	*сіз/ЖФ+дер/КЖ+де/ЖС*
---	----------	----------	----------
15	*сіздерден*	*сіз+дер+ден*	*сіз/ЖФ+дер/КЖ+ден/ШС*
16	*сіздермен*	*сіз+дер+мен*	*сіз/ЖФ+дер/КЖ+мен/КС*
17	*сіздерменен*	*сіз+дер+менен*	*сіз/ЖФ+дер/КЖ+менен/КС*
---	----------	----------	----------

3-блок. На основе сказанного в предыдущих 2-х блоках можно проследить образование всевозможных словоформ, которые также будут размеченными на морфологическом уровне, где будут указаны основа слов с указанием на часть речи, состав основы, а также морфологически размеченные словозменительные аффиксы.

Например, если за основу взять слово-существительное "***әке***", то данное слово в первоначальном поиске по списку (по базе) реестровых основ слов, начинающихся с начальных двух букв "***әк***", то мы найдем это слово с разметкой о принадлежности к "***зт.***", т.е. к существительному, и после этого согласно специальной компьютерной программе мы сможем образовывать всевозможные словоформы: ***әкедей, әкесіз, әкесіздің, әкесізге*** и т.д., которые по 4 столбцу ***таблицы 2*** будут размечены на грамматическом уровне, т.е. к словоизменительным аффиксам будут приписаны соответствующие условные обозначения в сокращенном варианте.

Обсуждение. Специалисты в этой области рассматривали корпусную лингвистику как одну из лингвистических областей, которые изучают ситуацию создания и использования языковых корпусов. Некоторые ученые считают эту тему узкой и объясняют ее только в области компьютерной лингвистики: "Корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с использованием компьютерных технологий" [10, с. 7]. А понятие компьютерной лингвистики, как правило, можно интерпретировать как широкий спектр использова-

ния компьютерных инструментов. Здесь мы называем «компьютерные инструменты» компьютерными программам и, обработкой языковых данных и правильной организацией компьютерных технологий и т. д. [11, с. 13-38].

А корпусная лингвистика использует компьютер только как “инструмент”. Вот почему корпусная лингвистика не может обойтись без компьютерных технологии. Однако, учитывая, что компьютер играет роль во всех формах современного образования, их всех нельзя отнести к компьютерной лингвистике.

Тексты представляют собой не только коллекцию электронных версий текстов на разных языках, но и современный языковой инструмент, основанный на компьютерной программе, которая автоматически анализирует тексты на языковые уровни. Поэтому необходимо создавать программы, которые делают такие автоматические анализы на корпусах. Учитывая тот факт, что компьютер способен работать с формальными моделями, необходимо ввести лингвистическое руководство по компьютерной программе. Поэтому лингвистическое развитие каждого уровня языка, позволяющее компьютеру автоматически проводить языковой анализ, является важной проблемой, стоящей перед лингвистами, которые готовят лингвистические продукты в прикладном направлении.

В казахском языкознании стоит вопрос создания национальных корпусов с ранних этапов XXI века. Отдел прикладной лингвистики института Языкознания имени А.Байтурсынова принимает активное участие в создании национальных корпусов в соответствии с требованиями информационных технологий. По инициативе профессора А.К. Жубанова методы и технологии корпуса были исследованы и изучены, под авторством А.Жубанова и А.Жанабековой был написан учебник “Корпусная лингвистика” [12, с. 5]. Практически с 2009 года были разработаны программы по внедрению лингвистических разметок и накоплен значительный опыт. В то же время впервые были выполнены анализы уровня морфологии. Эта автоматическая программа называется морфологическим анализатором. Программу создали программист института Языкознания имени А. Байтурсынова Д. Токмырзаев и бывший научный сотрудник института информатики К. Койбагаров. При поддержке Д. Токмырзаева и К. Койбагарова реализуются лингвистические и экстралингвистические замыслы, а также идеи математика и лингвиста, специалиста по прикладной лингвистике, профессора А. Жубанова и специалиста по прикладной лингвистике и грамматике, доктора филологических наук А.Жанабековой. Хорошо известно, что разработка лингвистических определений требует знаний в области лингвистики. Профессор А. Жунисбек занимается не только теоретическими проблемами фонетики, но и программными областями прикладной лингвистики - учебниками, методологией, синтезом и анализом речи. В то же время фонетические разметки корпуса основаны на трехступенчатом руководстве А. Жунисбека [13, с. 80], [14, с. 53-58]. В заключении, специалисты компьютерной лингвистики говорят, что фонд компьютерных языков – это способность ученого по-новому взглянуть на свой новый предмет. Чем больше лингвистическая основа, чем глубже языковая структура, тем глубже концепция исследуемого объекта и области человеческих знаний. Аналогичным образом значительно возрастут способности исследователя, появятся творческие источники энергии, и эти новые возможности, безусловно, будут использованы для улучшения системного характера казахского языка и его тщательного понимания.

Метатекстовая разметка – важная информация, которая должна быть включена в любом случае, и лингвистические разметки будут сделаны до введении текстов из разных стилей. Известно, что исследования будут проводиться по конкретной системе. Поскольку язык является очень обширной и большой системой средств общения, изучение языка может основываться на художественных или печатных или исторических памятниках, а также на конкретном жанре, стиле или конкретной теме, классифицируются как спирали.

Передача текстов с систематической структурой, включенных в память корпуса, возможена только при внедрении теоретических и практических методов метаразметки. Поэтому проблема метаразметки должна быть правильной, независимо от типа корпуса.

Метаразметки – в исследовательской работе. незаменимый источник для сбора материалов связанных с определением периода, стиля, автора, темы и т.д. Метаразметки дают возможность исследователю быстро найти информацию по стилю, периоду, автору и т.д. И это достижение, безусловно, является единственной силой для развития научно-исследовательского потенциала современной развитой страны.

Проблема метаразметок текстов корпуса изучается в отделе прикладной лингвистики института Языкознания им. А.Байтурсынова по грантовому проекту на 2015-2017 годы и по сей день. Также в 2016 году институт разработал метаразметку для целевого проекта под руководством А.Фазылжановой “Разработка и создание национального корпуса Казахского языка”. Специальное руководство было выпущено для метаразметчиков и каждый стиль основан на следующих пунктах, полях, в ячейках для типов вложения корпуса. Ячейка состоит из 23 пунктов. Во время работы по внедрению корпуса на основе латинского алфавита данная программа разработки метаразметки была продолжена. Таким образом, была создана полуавтоматическая программа ввода метаразметки.

Рассмотрим принцип морфологической разметки любой словоформы (или слова) очередного предложения из текстового корпуса казахского языка, согласно описаниям, представленным в предыдущих 3-х блоках.

Во-первых, допуская, что очередное слово предложения может быть не словоформой, а реестровым словом казахского языка, для ее разметки необходимо искать такое слово по ее начальным двум буквам в словарной базе основ казахских слов, где слова расположены строго по алфавиту, а также размечены по частям речи и по составу (односложное, двухсложное или многосложное). В случае совпадения проверяемое слово из предложения заменяется на размеченное слово из базы “Словаря основ казахского языка”.

Во-вторых, если в процессе поиска проверяемое слово из предложения не совпало ни с одним словом в размеченной базе “Словаря основ казахского языка”, то необходимо будет:

1) определить количество букв в данной словоформе;

2) по данному слову, в цикле образовывать словоформы по базе реестровых основ слов, с совпадением начальных двух букв и выбирать из порожденных словоформ те, которые совпадают также по “длине”, т.е. по количеству букв в словоформе. Допуская при этом, что это слово относится к существительному и воспользоваться базой “Слоизменительных аффиксов для существительных” для всех списков, учитывающих всевозможные варианты фонетических правил на стыке конечной буквы основ слов и словоизменительных аффиксов.

Данный процесс можно повторить, если допустить, что словоформа относится и к другим частям речи, воспользовшись базой (списком) словоизменительных аффиксов соответственно к части речи и конечной букве основ слов.

В-третьих, каждая вновь порожденная словоформа должна сравниваться с очередной проверяемой словоформой из предложения и в случае их совпадения порожденную словоформу можно считать грамматической размеченной, т.к. к основе слов уже будет приписана принадлежность ее к определенной части речи, а присоединенному к основе словоизменительному аффиксу также будет приписана грамматическая разметка, согласно его структуре и составу.

В-четвертых, для полуавтоматической разметки омономичных слов мы вносим такое предложение:

1) При сравнений очередного слова из предложения со словами из “Словаря основ казахского языка” (без учета разметки по частям речи) после первого совпадения сравниваемых единиц необходимо будет зафиксировать этот случай и повторить данный процесс сравнения до тех пор, пока не получим отрицательного ответа, т.е. “больше нет совпадений”. Это объясняется тем, что омонимичные слова могут повторяться несколько раз, причем, они могут принадлежать к разным частям речи или даже к одной и той же части речи, но в разных значениях. Например, омонимичные казахские слова: ***қара*** (***смотри***-глагол), ***қара*** (***черное***-прил.-ное), ***қара*** (***“силует”***-существ.) и т.д. Подобных омонимичных слов в казахском языке может быть в достаточной степени много [15, с. 113];

2) При сравнении очередной словоформы из предложения с вновь порожденными словоформами нет необходимости останавливаться при первом же совпадении из данных базы, например, со “Слоизменительными аффиксами для существительных” [16, с. 199], а необходимо зафиксировать данный случай и повторить данный процесс сравнения для баз со словоизменительными аффиксами, относящихся и к другим частям речи казахского языка до тех пор, пока не получим отрицательного ответа, т.е. “больше нет совпадений по другим частям речи”.

Заключение. В настоящее время в отделе Прикладной лингвистики Института языкознания имени А.Байтурсынова с этими 23 параметрами работает корпус текстов. Таким образом, необходимость введения метаразметок в создании корпуса не вызывает сомнений. Они позволяют пользователю корпуса, особенно лингвистам-исследователям и исследователям других областей науки быстро и легко найти любую необходимую информацию. Все вышесказанное вносится нами, конечно, только как предложение или просто идея, но, как нам кажется, вполне осуществимая.

В заключении хотелось бы сказать, что в настоящее время мы, в основном, рукодствуемся методами автоматизации мофологической разметки текстов на русском, украинском и некоторых тюркских языках [17], [18], [19] Использование данного принципа для автоматической разметки (или полуавтоматической разметки) казахских текстов на практике показало не плохие результаты, хотя, конечно, некоторые моменты требуют совершенствования и практической доработки.

Список использованной литературы:

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. – С 32. – Книга

2 Исаев С. Қазақ тіл білімінің мәселелері. – Алматы: Арыс, 2008. – 624 б. – Книга

3 Плунгян В. А. Общая морфология: Введение в проблематику. – М.: Эдиториал УРСС, 2000. – Книга

4 Зализняк А.А. Грамматический словарь русского языка. Словоизменение. – М., 1980. – 880 с. – Словарь

5 Шаяхметов Қ. Екі функциялық аффикстер: филол. ғыл. канд. дисс. – Алматы, 1973. – Диссертация

6 Насилов В.М. “Аффиксы включения” //Сб. Вопросы языка и литературы стран востока. – М. 1958. – Сборник

7 Ганиев Ф.А. “О синтетических и аналитических падежах в татарском языке.” // Сб.Вопросы тюркологии. – Казань, 1970. – Сборник

8 Хабичев М.А. Именное словообразование и формообразование куманских языках. – М.:Наука, 1989. – 217 с. – Книга

9 Баскаков Н.А. Историко-типологическая морфология тюркских языков. – М.:Наука, 1979. – Книга

10 Захаров В.П., Богданова С.Ю. Корпусная лингвистика. Иркутск: ИГЛУ, 2011. – С. 7. – Учебное пособие

11 Баранов А.Н. Компьютерная лингвистика // Введение в прикладную лингвистику: Учебное пособие. – М.: Едиториал УРСС, 2003. – С. 13-38. – Учебное пособие

12 Жубанов А.К., Жанабекова А.А. Корпусная лингвистика. – Алматы: Казак тили, 2017. – С 5. – Учебное пособие

13 Жунисбек А. Проблемы казахского языкознания. – Алматы: Абзал-ай, 2018. – С. 80. – Книга

14 Жаңабекова А. Мәтіндер корпусына фонетикалық белгіленімдер енгізуге арналған нұсқаулық сипаттамасы // Ә. Жүнісбектің 80 – жылдық мерейтойына арналған "Қазақ фонетикасы" атты дөңгелек үстел материалдары. – Алматы, 2018., 19 қазан. – 53 - 58 бб. – Материалы научной конференции

15 Исаев С. Қазіргі қазақ тіліндегі сөздердің грамматикалық сипаты. – Алматы: Рауан, 1998. – 303 б. – Книга

16 Ысқақов А. Қазіргі қазақ тілі. – Алматы, 1991. – 382 б. – Учебное пособие

17 Азарова И.В. Морфологическая разметка текстов на русском языке с использованием формальной грамматики AGFL. Кафедра математической лингвистики СПбГУ // http://www.dialog-21.ru/Archive/2003/AzarovaAGFL. htm. –Интернет ресурс

18 Национальный корпус русского языка // http://www.ruscorpora.ru/ – Интернет ресурс

19 Дарчук Н.П. Автоматизований морфологічний аналіз тексту // http://linguist.univ.kiev.ua/courses_morph.htm – Интернет ресурс

References:

1. Sirazitdinov Z.A., Buskunbaeva L.A., İşmuhametova A.Ş., İbragimova A.D. İnformasionnye sistemy i bazy dannyh başkirskogo iazyka. – Ufa: Knijnaia palata RB, 2013. – S 32. – Kniga
2. İsaev S. Qazaq tıl bılımınıñ mäseleleri. – Almaty: Arys, 2008. – 624 b. – Kniga
3. Plungän V. A. Obşaia morfologia: Vvedenie v problematiku. – M.: Editorial URSS, 2000. – Kniga
4. Zaliznäk A.A. Gramaticheski slovär ruskogo iazyka. Slovoizmenenie. – M., 1980. – 880 s. – Slovär
5. Şaiahmetov Q. Ekı funksialyq afikster: filol. ğyl. kand. diss. – Almaty, 1973. – Disertasia
6. Nasilov V.M. "Afiksy vklüchenia" //Sb. Voprosy iazyka i literatury stran vostoka. – M. 1958. – Sbornik
7. Ganiev F.A. "O sinteticheskih i analiticheskih padejah v tatarskom iazyke." // Sb.Voprosy türkologii. – Kazän, 1970. – Sbornik
8. Habichev M.A. İmenoe slovoobrazovanie i formoobrazovanie kumanskih iazykah. – M.:Nauka, 1989. – 217 s. – Kniga
9. Baskakov N.A. İstoriko-tipologicheskaia morfologia türkskih iazykov. – M.:Nauka, 1979. – Kniga
10. Zaharov V.P., Bogdanova S.İu. Korpusnaia lingvistika. İrkutsk: İGLU, 2011. – S. 7. – Uchebnoe posobie
11. Baranov A.N. Kömpüternaia lingvistika // Vvedenie v prikladnuiu lingvistiku: Uchebnoe posobie. – M.: Editorial URSS, 2003. – S. 13-38. – Uchebnoe posobie 12
12. Jubanov A.K., Janabekova A.A. Korpusnaia lingvistika. – Almaty: Kazak tili, 2017. – S 5. – Uchebnoe posobie
13. Junisbek A. Problemy kazahskogo iazykoznania. – Almaty: Abzal-ai, 2018. – S. 80. – Kniga
14. Jañabekova A. Mätınder korpusyna fonetikalyq belgılenımder engızuge arnalğan nūsqaulyq sipattamasy // Ä. Jünısbektıñ 80 – jyldyq mereitoiyna arnalğan "Qazaq fonetikasy" atty döñgelek üstel materialdary. – Almaty, 2018., 19 qazan. – 53 - 58 bb. – Materialy nauchnoi konferensii
15. İsaev S. Qazırgı qazaq tılındegı sözderdıñ gramatikalyq sipaty. – Almaty: Rauan, 1998. – 303 b. – Kniga
16. Ysqaqov A. Qazırgı qazaq tılı. – Almaty, 1991. – 382 b. – Uchebnoe posobie
17. Azarova İ.V. Morfologicheskaia razmetka tekstov na ruskom iazyke s ispölzovaniem formälnoi gramatiki AGFL. Kafedra matematicheskoi lingvistiki SPbGU // http://www.dialog-21.ru/Archive/2003/AzarovaAGFL. htm. – İnternet resurs
18. Nasionälnyi korpus ruskogo iazyka // http://www.ruscorpora.ru/ – İnternet resurs
19. Darchuk N.P. Avtomatizovani morfologıchni analız tekstu // http://linguist.univ.kiev.ua/courses_morph.htm – İnternet resurs

МРНТИ 17.17.01 https://doi.org/10.51889/2020-4.1728-7804.61

Raeva G.,[1] Ilyassova N.[2]

[1,2]Kazakh national pedagogical university named after Abai ,
Almaty, Kazakhstan

TURKIC WORD: VARIANTS AND MEANING

Abstract

The article deals with variant words in the Turkic languages. From the historical point of view the significance of synharmonic variants and parallels is in the following: the phenomenon plays a special role in a methodological aspect in order to restore ancient roots and phonemes in their previous forms. We vividly observe the fact when we compare words