УДК 81-13:811.512.154

ЧАСТЕРЕЧНЫЕ РАЗМЕТКИ ДЛЯ НОВОГО КОРПУСА КЫРГЫЗСКОГО ЯЗЫКА

(Инструментарий Turkic Lexicon Apertium)

А.А. Касиева, А.Т. Сатыбекова

Корпусная лингвистика является новым направлением в кыргызской лингвистике и представляет собой малоизученную парадигму, а исследования в этой области на данный момент весьма скудны. Главной целью данной статьи является ознакомление с процессом тэггирования, другими словами, частеречной разметкой текстов новосозданного корпуса кыргызского языка. Хотя сам процесс является в достаточной степени трудоемким и времязатратным, требуется индивидуальный подход для решения каждого из ряда задач, существующих в лингвистике. Поскольку кыргызский язык относится к агглютинативным языкам, соответственно, данный процесс усложняется вдвойне. После многократных обсуждений для морфологической разметки словоформ корпуса было решено использовать инструментарий Turkic Lexicon Apertium – платформу машинного перевода. В практической части данной статьи представлен подробный анализ предложений на морфологическом и синтаксическом уровнях, которые были извлечены из новосозданного корпуса кыргызского языка. Результаты исследования будут использованы для дальнейшего развития корпуса кыргызского языка, вызовут интерес у студентов, магистрантов и лингвистов, а также послужат мотивацией для их вовлечения в этот увлекательный процесс построения корпусов разных типов.

Ключевые слова: корпус кыргызского языка; тэггирование; корпусная лингвистика; частеречная разметка; морфологический анализ; Turkic Lexicon Apertium.

КЫРГЫЗ ТИЛИНИН ЖАҢЫ КОРПУСУНДАГЫ СӨЗ ТҮРКҮМДӨРДҮН ЭНТЕКТЕРИ

(Turkic Lexicon Apertium аспаптарына ылайыкташтыруу)

А.А. Касиева, А.Т. Сатыбекова

Корпустук лингвистика кыргыз тил илиминде жаңы жана аз изилденген тил парадигмасы болуп саналат, ал эми бул тармакка байланыштуу илимий эмгектердин саны азыркы учурда жокко эсе. Бул макаланын негизги максаты – жаңы түзүлгөн кыргыз тилинин корпусунда жайгаштырылган тексттердеги сөз формаларды морфологиялык энтектер менен белгилөө, башкача айтканда, аларга тэг (англ. ‘tag’) ыйгаруу процесси менен тааныштыруу болуп саналат. Бул процесс көп убакыт жана аракетти талап кылгандыктан, ар кандай тилдик маселени чечүүдө ага жараша чечим кабыл алынат. Мындан улам корпустагы сөздөргө энтек ыйгаруу көбүнэсе кол менен иштелип чыгат. Ал эми кыргыз тили табияты боюнча агглютинативдик тил болгондугун эске ала турган болсок, бул жагдай мындан да татаал абалга туш болот. Бул макалада көтөрүлгөн маселени чечүүдө көптөгөн талкуулар орун алды жана алардын натыйжасында корпустун материалдарын морфологиялык жактан белгилөө үчүн Turkic Lexicon Apertium – машиналык котормо платформасынын стандарттык аспаптарын колдонуу чечими кабыл алынды. Макаланын практикалык тарабы жаңы кыргыз корпусунун сүйлөмдөрүнө морфологиялык жана синтаксистик деңгээлде жүргүзүлгөн деталдуу анализин камтыйт. Бул макаланын жыйынтыктары кыргыз корпусунун өнүгүшүнө салым кошуп, студенттердин, магистранттардын жана тилчилердин кызыгуусун арттырат деп ишенебиз.

Түйүндүү сөздөр: кыргыз тилинин корпусу; энтек-белгилөө; корпустук лингвистика; сөз түркүмдөрдүн белгилениши; морфологиялык талдоо; Turkic Lexicon Apertium.

PARTS-OF-SPEECH ANNOTATION OF THE NEWLY CREATED KYRGYZ CORPUS

(Turkic Lexicon Apertium Tools)

A.A. Kasieva, A.T. Satybekova

Corpus linguistics is a new direction in Kyrgyz linguistics and is a little-studied paradigm, and research in this area is currently very scarce. The main goal of the paper is to get acquainted with the process of part-of-speech tagging of word-forms of the Kyrgyz corpus. This paper is focused on the process of labelling the tokens of the newly created

Kyrgyz language corpus. Though the process itself is labour and time-consuming, it is usually performed manually. This procedure becomes even more complicated due to the agglutination of the Kyrgyz language. After long discussions it was decided to exploit standard toolkits of the Turkic Lexicon Apertium, an open-source machine translation platform. For annotating, each word is labelled and analyzed along with the process of tagging in closer observation. The section of discussion includes a detailed analysis of Kyrgyz sentences extracted from the Kyrgyz corpus. Morphological and syntactic analyses of the sentences are presented as samples. We believe that this work will give impetus for further development and enrichment of the Kyrgyz corpus and attract students and linguists to get involved in this interesting process.

Keywords: Kyrgyz corpus; POS-tagging; corpus linguistics; morphological analysis; Turkic Lexicon Apertium.

Введение. Термины “корпус” и “корпусная лингвистика” являются не столь новыми для мирового научного сообщества XXI века, но можно утверждать, что они являются одними из важных составляющих научно-технического прогресса, так как наитеснейшим образом взаимосвязаны с развитием технологий, будучи частью прикладной и компьютерной лингвистики. История развития корпусной лингвистики берёт своё начало с середины XX века, точнее с 1961 года, когда в США был создан первый Брауновский корпус американского варианта английского языка. Авторами корпуса, который состоял из одного миллиона слов (500 текстов по 2000 слов в каждом), были У. Фрэнсис и Г. Кучера [1]. Корпус привлек внимание и глубокий интерес научного сообщества того времени, в связи с чем наряду с общественным резонансом начали вестись различные оживленные дискуссии. Поначалу было также очень много критики. Ярым представителем отрицания нового веяния того времени, можно сказать, был Н. Хомский. Как отмечают Т. МакЭнери и Э. Вилсон в своей книге *Corpus Linguistics* [2], рассуждения Хомского относительно корпусной лингвистики, вызвавшие бурю споров среди лингвистов того времени, по своей сути, не являются новыми, но представляют спор между рационалистами и эмпиристами. Хомский изменил предмет лингвистического исследования от абстрактного описания языка к теориям, отражающим психологическую реальность, являющимся разумными и правдоподобными моделями языка. Таким образом, он отказался признавать “корпус” как достоверный источник для лингвистических исследований. Можно предположить, что такого же мнения он придерживается и в эти дни, согласно его высказываниям, озвученным во время симпозиума “Brains, Minds and Machines” в честь 150-летней годовщины Массачусетского технологического института в 2011 году. Хомский выразил свое мнение относительно точности подходов, применяемых в сфере развития искусственного интеллекта. Он считает, что статистический подход может иметь практическую ценность для поисковых систем при наличии компьютеров, которые способны обрабатывать огромные массивы данных, что в свою очередь противоречит научным взглядам, в которых данный подход является более поверхностным и менее адекватным.

В 70-х годах XX века, немного позднее с момента появления Брауновского корпуса в США, был создан корпус и для британского варианта английского языка под названием “Ланкастер-Осло-Берген” (ЛОБ).

Следует заметить, что до появления компьютеров текстовые материалы хранились в библиотеках, поэтому корпуса, по своей сути, не являются самыми первыми репозиторами массивов текстов. Все работы, связанные с материалами библиотек докомпьютерной эры, велись исключительно на бумажной основе. Так, например, ведение картотеки было не особо удобным и времязатратным по сравнению с тем, что на сегодняшний день может корпус. Так, относительно в недавнем прошлом начали появляться корпуса для разных языков мира. В связи с тем что само понятие “корпусная лингвистика” является практически необъятным, до сих пор еще нет точного универсального определения для понятия “корпус”.

Что такое “корпус” и “корпусная лингвистика”? На сегодняшний день существует множество определений понятию “корпус”, которое в переводе с латинского означает “тело, единое целое”. В Оксфордском словаре дается следующее определение: “Corpus is a collection of written or spoken texts” [3]. Разные ученые дают разные определения данному понятию. Так, Дж. Лич определяет корпус как коллекцию данных явлений естественного языка, который

может послужить основой для лингвистических исследований и включающий в себя как письменные тексты, так и транскрипции записи устной речи, созданные для представления определенного языка или языковых вариантов [4]. По мнению Э. Финегана, корпус представляет собой репрезентативное собрание текстов в машиночитаемом формате, включающее информацию о ситуации, в которой текст был произведен [1, с. 11].

Споры касательно направленности корпусной лингвистики, точнее говоря, следует ли рассматривать корпусную лингвистику как отдельный раздел языкознания или же только как методологию в языкознании, актуальны до сих пор. Это связано с двойственностью характера объекта корпусной лингвистики. То есть объект корпусной лингвистики может выступать как исходный материал для других исследований, так и его продуктом. Междисциплинарная ориентированность корпусной лингвистики дает возможность рассматривать и решать лингвистические задачи на разных его уровнях. Поэтому в настоящее время наблюдается тенденция исследования языковых задач посредством корпусных данных. Также практическую ценность корпусная лингвистика представляет и для переводчиков, лексикографов, исследователей компьютерной лингвистики, а также многих других исследователей широкого круга науки.

При создании корпусов одними из принципиальных и важных вопросов являются тип и объем текстов, которые отбираются в соответствии с требованиями и критериями корпуса. К наиболее важным критериям корпуса можно отнести: вид текста (письменный, устный, электронный); тип (книга, журнал, газета, статья и т. д.); характер текста (научный, публицистический, официальный, художественный и т.д.); используемый язык/подъязык/диалект; место и периоды появления текстов. Совершенно естественно, что тексты не попадают в корпус случайно, кроме вышеперечисленных критериев, они также должны отображать проблемную область исследований языковых явлений. Проблемная область имеет два аспекта: языковой и речевой. Языковой аспект – это само изучаемое явление, а речевой – то множество контекстов, в которых это явление представлено [5]. Так как проблемная область исследования может быть весьма широкой, соответственно, перед создателями корпусов стоит задача включить как можно большее количество текстов, которые относятся к определенному языку или его подмножеству, для изучения которого и создается корпус.

Для решения множества различных лингвистических задач необходимое условие, чтобы корпусные тексты содержали лингвистическую и металингвистическую информацию – разметку или же аннотацию, соответствующую разным уровням лингвистического описания, – фонетическая, морфологическая, синтаксическая, семантическая и прочие. При определении аннотации для конкретного лингвистического уровня необходимо учитывать цели, для достижения которых используются корпусные данные. Таким образом, следуя цели данной статьи, нами предпринята попытка описания и обсуждения морфологической разметки словоформ новосозданного корпуса кыргызского языка посредством использования стандартных инструментов Turkic Lexicon Apertium – платформы машинного перевода.

Морфологическая разметка для корпуса кыргызского языка. Корпус кыргызского языка на данный момент содержит в себе 1 205 888 слов, включенных из 84 письменных текстов разных жанров (эпосы, романы, рассказы, сказки и др.). Корпус кыргызского языка – это совместный проект между Кыргызско-Турецким университетом “Манас” и Университетом Саарланд, осуществленный под руководством профессора Э. Тайк [6].

Следует отметить, что вопрос частеречной разметки словоформ кыргызского языка ранее был исследован в статье Т. Садыкова [7], из которого мы и заимствовали термин “энтек”, т. е. “разметка”, “знак”. Данная статья является уникальной и своеобразной в кыргызской лингвистике, так как в ней в деталях перечислены морфологические разметки кыргызского корпуса, соответствующие Лейпцигскому стандарту. Тем самым эта работа вносит свой вклад в развитие корпусной лингвистики кыргызского языка.

В практической части статьи мы представляем образцы частеречной разметки предложения, извлеченные из тела корпуса [8]. Рисунок 1 представляет собой образец морфологической разметки предложения, извлеченного из корпуса кыргызского языка. Словоформы в предложении

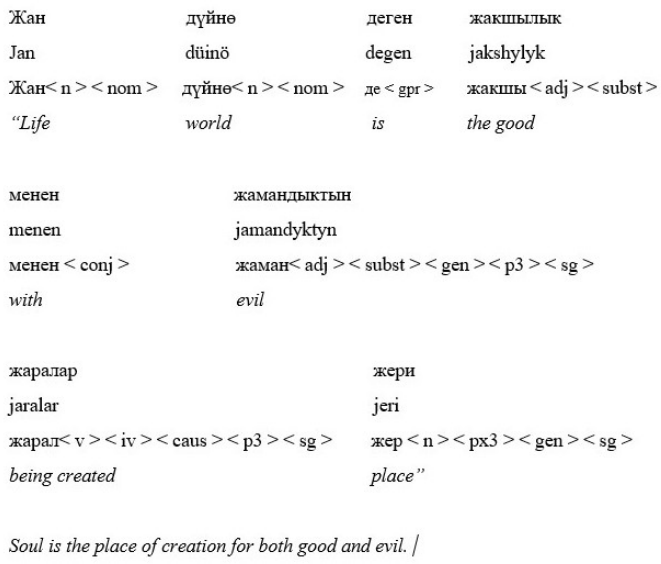

Рисунок 1. Образец морфологической разметки корпуса

проаннотированы тэггами (символами) Turkic Lexicon Apertium, протранскрибированы латинской графикой и переведены на английский язык.

1. Жан дүйнө деген жакшылык менен жамандыктын жаралар жери.

/$/

ˆ Жан дүйнө /жан дүйнө < з.ат. > < абстр. > < адамз. э. > < жек. с. > < ат жөн.>

ˆдеген/ < атоочт. >

ˆжакшылык/жакшы < сын ат. > < сап. с.ат > < жай дар. > +лык< субст. з.ат. > < абстр. > < адамз. э. > < жек.с. > < ат жөн.>

ˆменен/ < байл. >

ˆжамандыктын/жаман < сын ат. > < сап. с.ат > + дык< субс. з.ат. > < абстр. > < адамз. э. > < жек. с. > + тын< ил. жөн. >

ˆжаралар/жарал < эт. > < өтп. эт. > < туб. эт > +ар< ар. мам. > < III ж. > < жек. с. >

ˆжери/жер < з.ат. > < абстр. > < адамз. э. > < жек. с. > + и< III ж.таанд.м. > < ил. жөн.>

/$./

/$/

ˆ Жан дүйнө/жан дүйнө < n > < nom >

ˆдеген/деген < gpr >

ˆжакшылык/жакшы < adj > + лык < adj > < subst >

ˆменен/менен < conj >

ˆжамандыктын/жаман < adj > + дык< adj > < subst > +тын < gen > < p3 > < sg >

ˆжаралар/жарал < v > < iv >+ар< caus > < p3 > < sg >

ˆжери/жер < n > +и< px3 > < gen > < sg >

/$./

Жан дүйнө деген жакшылык менен жамандыктын жаралар жери.

Jan düinö degen jakshylyk menen jamandyktyn jaralar jeri.

*(*lit.: *"life world is the good with evil's creation place")*

Soul is the place of creation for both good and evil.

Далее мы предлагаем ознакомиться с техникой запроса информации по корпусу кыргызского языка. На рисунке 2 видим, какие символы необходимы для ввода искомого слова. Например, допустим, нам нужно извлечь информацию о слове ***"дүйнө"***, о статистических данных этого слова, плотности его встречаемости в кыргызском языке, в каком дискурсе наиболее часто используется, с какими словами чаще сочетается, встречается ли слово в начале предложения, в середине или в конце. Метадата корпуса также обеспечивает полной информацией о том, из какого источника извлечено это слово наряду с его морфологической разметкой.

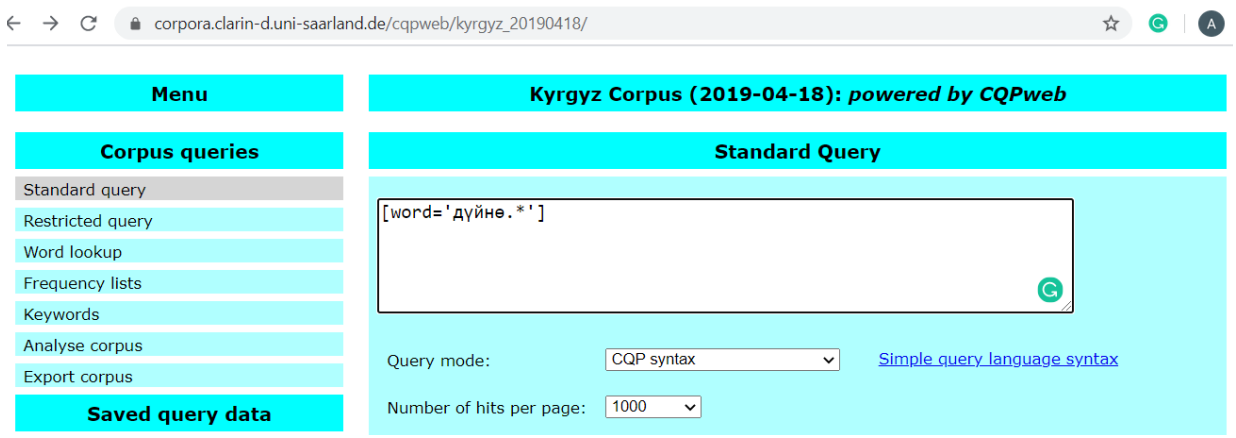

Рисунок 2. Образец ввода искомого слова в корпусе

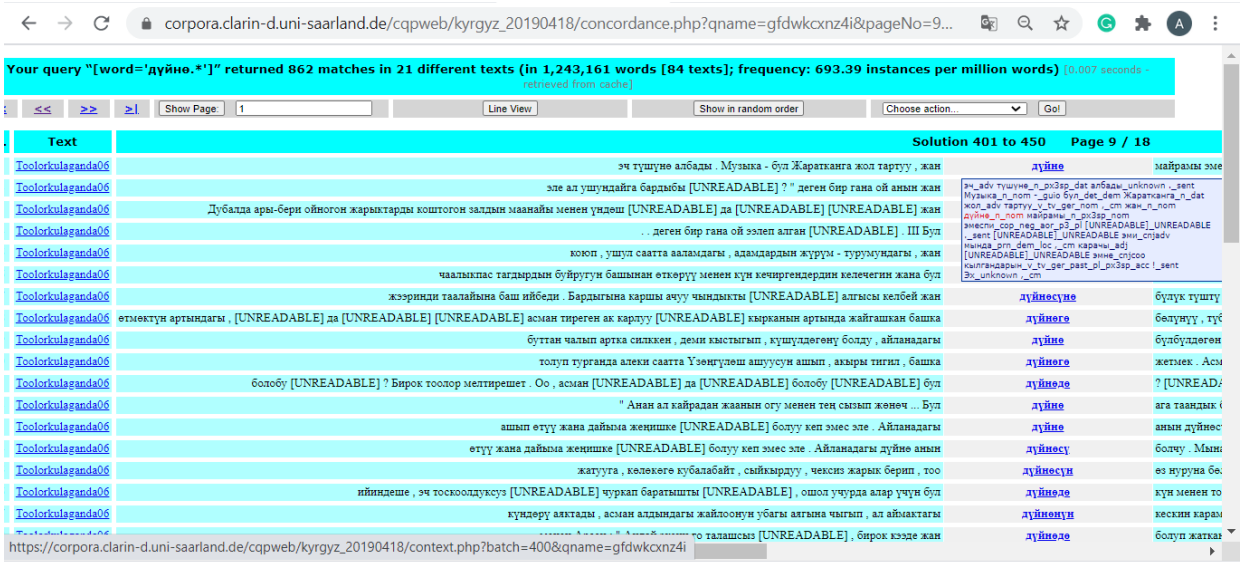

Рисунок 3. Результат запроса слова ***“дүйнө”***

Из рисунка 3 можно увидеть вывод данных, запрошенных по поиску слова ***“дүйнө”***, который показал, что слово ***“дүйнө”*** встречается 826 раз в 21 тексте корпуса, а частотность его встречаемости и распределенности по всему корпусу составляет 693.39 на миллион слов. Кроме этого, при наведении курсора на искомое слово мы видим, что слово ***“дүйнө”*** является существительным в именительном падеже.

Заключение. Направление корпусной лингвистики в кыргызском языке является на данный момент новым и пока еще находится на стадии проработки. Процесс тэггирования в кыргызском языке (“наслаивания” частеречной разметкой каждого токена с учетом контекста) является очень сложным в связи с тем, что не всегда и не во всех случаях применение одной разметки будет уместным. Также это сопряжено с тем, что процесс унификации еще не произведен, следовательно, предстоит проделать большую работу как над проработкой частеречных разметок и повышением его уровней, так и над увеличением корпуса кыргызского языка. Еще одной проблемой, с которой мы столкнулись во время морфологического разбора, – это проблема составных словосочетаний, глаголов и особенно фразеологизмов.

В силу агглютинативной природы кыргызского языка часть речи одного и того же слова не всегда остается той же в разных речевых ситуациях и контекстах, что и создает различного рода проблемы. Именно для решения таких проблем возникает необходимость унификации морфологических разметок для их дальнейшего использования в целях разметки данных корпуса кыргызского языка.

За оказанную поддержку при аннотировании предложений корпуса выражаем благодарность старшему преподавателю кыргызско-турецкой программы переводческого отделения КТУ "Манас" Дөөлөтбеку Эшкенову.

Литература

1. *Захаров П.* Корпусная лингвистика: учебник для студентов направления "Лингвистика" / П. Захаров, С. Богданова. СПб.: СПбГУ, 2013. С. 11.
2. *McEnery T.* Corpus linguistics. An introduction! / T. McEnery, A. Wilson // Book – second edition. Edinburgh: Edinburgh University Press, 1996. 6 p.
3. *Hornby A.S.* Oxford Advanced Learner's Dictionary: Oxford Advanced Learner's Dictionary of Current English – eighth edition / A.S. Hornby. Oxford University Press, 2010. 339 p.
4. *Garside R.* Corpus Annotation: Linguistic information from Computer Text Corpora / R. Garside, G. Leech, T. McEnery // Book. New-York: Routledge, 2013. 1 p.
5. *Янполова А.* Основные принципы построения лингвистических корпусов / А. Янполова / Пятигорский государственный университет. Пятигорск: Издания ПГУ Молодая наука, 2017. Ч. 11.
6. Conference materials "The book of abstract online": A new Kyrgyz corpus: sampling, compilation, annotation / A. Kasieva, J. Knappen, S. Fischer, E. Teich. Hamburg: De Gruytor, 2019. 316 p.
7. *Садыков Т.* Морфологические разметки для национального корпуса / Т. Садыков, Б. Шаршембаев, Б. Көчкөнбаева // Вестник КРСУ. 2018. Т. 18. № 1. С. 91–94.
8. URL: http: //corpora.clarin-d.uni-saarland.de/cqpweb/kyrgyz_20190418 (дата обращения: 02.06.2020).