УДК: 811.512.154

Мусажанова С. Ж., магистрант
musazhanova.seide@gmail.com
Касиева А. А., канд. филол. наук, доцент
aida.kasieva@manas.edu.kg
Джумалиева Г. К., докт. филол. наук, доцент,
gulnur.jumalieva@manas.edu.kg
КТУ "Манас", Кыргызстан

СИНТАКСИЧЕСКАЯ АННОТАЦИЯ КЫРГЫЗСКОГО ЯЗЫКА НА ОСНОВЕ НОВОСОЗДАННОГО КОРПУСА

В статье рассматриваются вопросы, связанные с областью корпусной лингвистики, а в частности, она посвящается проблемам синтаксического аннотирования (разметки) предложений в кыргызском языке, взятых из кыргызского корпуса. Особое внимание при этом уделяется грамматическим особенностям рассматриваемого языка и проблемам, возникающим с их синтаксическим аннотированием с использованием веб-приложений платформы Универсальных Зависимостей (УЗ) (Universal Dependencies), который предназначен для автоматизированной обработки текста: синтаксис и грамматика. Схема аннотации УЗ кыргызских предложений производится в виде деревьев зависимостей. В результате исследования выявлено, что не все грамматические категории кыргызского языка достаточно изучены в плане их рассмотрения в рамках УЗ. Следовательно, предстоит проделать немало работ для усовершенствования кыргызского корпуса и корпус послужит надежной базой для дальнейших исследований в области компьютерной корпусной лингвистики.

Ключевые слова: *корпусная лингвистика, кыргызский язык, корпус кыргызского языка, синтаксическая аннотация, части речи, универсальные зависимости, древовидные зависимости, парсинг.*

Мусажанова С. Ж., магистрант
musazhanova.seide@gmail.com
Касиева А. А., филол. илим. канд., доцент
aida.kasieva@manas.edu.kg
Джумалиева Г. К., филол. илим. докт., доцент
gulnur.jumalieva@manas.edu.kg
"Манас" КТУ, Кыргызстан

КЫРГЫЗ ТИЛИНИН ЖАҢЫ ТҮЗҮЛГӨН КОРПУСУНУН СИНТАКСИСТИК ЭНТЕКТЕЛИШИ

Макалада корпустук лингвистика тармагына байланыштуу маселелер каралат, атап айтканда, Кыргыз корпусунан алынган кыргыз тилиндеги сүйлөмдөрдү синтаксистик аннотациялоо (энтектөө) маселелерине арналат. Изилдөөдө кыргыз тилинин грамматикалык өзгөчөлүктөрүнө жана текстти автоматташтырылган түрдө иштеп чыгууга арналган универсалдуу көз карандылык платформасынын (UD) веб-тиркемелерин колдонуу менен алардын синтаксистик аннотациясында пайда болгон көйгөйлөргө

(синтаксис жана грамматика) өзгөчө көңүл бурулат. Кыргызча сүйлөмдөрдүн аннотацияланышынын схемасы көз карандылык дарагы түрүндө жүргүзүлөт. Изилдөөнүн жыйынтыгында кыргыз тилинин бардык грамматикалык категориялары жетиштүү деңгээлде изилдене электиги аныкталды. Демек, кыргыз корпусун өркүндөтүү үчүн көп иштерди аткаруу милдети турат жана андан соң гана кыргыз корпусу мындан аркы изилдөөлөр үчүн ишенимдүү база болуп кызмат кыла алат.

***Өзөктүү сөздөр**: корпустук лингвистика, кыргыз тили, кыргыз тилинин корпусу, синтаксистик аннотация, сөз түркүмдөрү, универсалдык көз карандылык, дарак түрүндөгү көз карандылык, парсинг.*

Musazhanova S., master's student,
musazhanova.seide@gmail.com.
Assoc.Prof. Dr. Kasieva A., Department of Simultaneous Translation,
aida.kasieva@manas.ed.kg.
Assoc.Prof. Dr.of Philology Dzhumalieva G.,
gulnur.jumalieva@manas.ed.kg
Kyrgyz-Turkish Manas University, Kyrgyzstan

SYNTACTIC ANNOTATION OF THE NEWLY-CREATED KYRGYZ CORPUS

The article discusses the issues related to the field of corpus linguistics and, particularly, focuses on the parsing (annotation) of sentences in the Kyrgyz language taken from the Kyrgyz corpus. A special interest is paid to the grammatical features of the language in question and the problems encountered in their syntactic annotation with the use of the Universal Dependencies (UD) web platform, which is designed for automated text processing: syntax and grammar. The annotation scheme of UD of the Kyrgyz sentences is generated in the form of dependency trees. As a result of the study, it was revealed that not all grammatical categories of the Kyrgyz language are sufficiently studied in terms of their consideration within the framework of UD. Therefore, much work remains to be done to improve the Kyrgyz corpus and it will serve as a reliable source for further research in the field of computer corpus linguistics.

***Keywords**: Corpus Linguistics, Kyrgyz language, Corpus of the Kyrgyz language, syntactic annotation, parts of speech, Universal Dependencies, tree dependencies, parsing.*

Введение. В наши дни технологии проникли во все сферы жизнедеятельности человека. Это стало одним из самых важных аспектов нашей жизни, который теперь влияет на использование человеческого языка. Особенно это касается использования языка через гаджеты, таких как компьютеры, смартфоны, разные веб-приложения и т.д. Технология открыла огромные возможности, появились программные обеспечения на базе множества языков. Таким образом, компьютерная обработка естественного языка, которая является одной из ветвей искусственного интеллекта, стала неотъемлемой частью лингвистической науки.

Современная лингвистика и корпусные исследования становятся более комплексными и интегральными чем когда-либо. Появилась

возможность изучать все уровни языка, используя разного рода веб-приложения и лингвистические базы данных. Например, можно изучать фонетику, морфологию, синтаксис, семантику и прагматику в рамках отдельно взятого лингвистического корпуса или нескольких корпусов. Точно также, язык выходит за рамки чисто лингвистических границ, соприкасаясь и с другими дисциплинами таких как социолингвистика, психолингвистика, теоретическая/прикладная лингвистика, географическая лингвистика и другие. [1]

Соответственно, все вышеперечисленные процессы предусматривают рассмотрение специальных и эффективных методов для их исследования. Так, например, на стыке синтаксиса и корпусной лингвистики появились новые методы морфологических и синтаксических аннотаций и/или разметок для слов или целых предложений. Их можно совершать на разных платформах в онлайн режиме и параллельно работать с корпусами.

Корпусная лингвистика - это изучение языка на основе письменных текстов любого языка. Корпусная лингвистика предполагает, что анализ реального языка более возможен с корпусами (базами данных), собранными в полевых условиях в их естественном контексте [2]. Это, безусловно, сильно отличается от других тем в лингвистике, поскольку непосредственно не связано с изучением какого-либо конкретного аспекта языка. Скорее, это область, которая фокусируется на наборе процедур или методов для изучения языка. Корпусную лингвистику можно определить, как набор машиночитаемых текстов и их анализ. Она обладает потенциалом для исследования и изменения всего нашего традиционного подхода к изучению языка. Самое главное, что развитие корпусной лингвистики также привело к исследованию новых теорий и ряда старых, которые теперь облегчают исследование данных теорий непосредственно посредством языковых корпусов.

Использование кыргызского языка в социальных сетях и интернете в данное время можно характеризовать ограниченным ввиду того, что гаджеты не всегда могут распознавать голосовой ввод на кыргызском языке. Это и является основной причиной того, что сеть не способна обрабатывать данные просто потому, что они еще не полностью компьютеризированы («системы не обучены») и аннотированы (разметка частей речи). Следовательно, обработка естественного языка, создание корпусов и проведение морфологических и синтаксических аннотаций также должны применяться на кыргызском языке, чтобы в дальнейшем анализировать, переводить и исследовать тексты, изучать языки и многое другое.

Основной целью данной статьи является ознакомление и введение в кыргызскую лингвистику ряда новых теорий и терминов, связанных с этой областью, а также апробация специальных веб-приложений для синтаксического разбора кыргызских предложений. Точнее говоря, впервые осуществляется попытка синтаксического разбора кыргызского языка, используя универсальные зависимости (парсеры) и древовидных зависимостей.

Корпусная лингвистика, машинный перевод и обработка естественного языка - это новые термины для кыргызского языка. Хотя машинный перевод и корпусная лингвистика хорошо известны в других странах мира, они являются совершенно новыми понятиями в нашем языке и обществе, которые сейчас мы связываем с недавним созданием корпуса кыргызского языка (18/04/2019) и (08/03/2022), который содержит всего более 2 миллионов слов. Это количество намного меньше и невелико по сравнению с другими уже обработанными языками, которые могут содержать более миллиарда слов: British National Corpus (BNC), Corpus of Contemporary American English (COCA), Chinese Corpus (CC), Национальный корпус русского языка и многие другие. Самые популярные языки, на котором говорят миллионы людей, были уже аннотированы (морфологически, синтаксически). К примеру, национальный корпус русского языка содержит более 2 миллиардов слов из разнообразных текстов, которые были успешно оцифрованы [3]. Следовательно, данное исследование призвано внести вклад в ново созданный корпус кыргызского языка в плане его синтаксической разметки и в целом в его развитие. Для достижения цели данного исследования мы представляем морфологическую и синтаксическую аннотации предложений из кыргызского корпуса, используя международную платформу Универсальных Зависимостей (УЗ).

Первый национальный корпус кыргызского языка. На базе Кыргызско-Турецкого университета «Манас» и Саарландского университета была начата работа над корпусом кыргызского языка, который был создан в апреле 2019 года [4]. Кыргызский корпус невелик и находится на стадии развития; он включает в себя 2 миллиона слов из текстов художественной литературы и средств массовой информации (газета «Эркин-Тоо») (см. Рисунок 1) [5].

Примерно 4,4 миллиона человек во всем мире говорят на кыргызском языке, который является государственным языком в Кыргызстане. Он принадлежит к тюркской языковой семье и имеет богатую агглютинативную морфологию. Кыргызский в настоящее время является

языком с ограниченными ресурсами в плане корпусной лингвистики; имеются также другие доступные кыргызские веб-корпуса, но без аннотаций. По этой причине, нынешний корпус является первым кыргызским корпусом, содержащим морфологическую аннотацию, которая предварительно была сделана вручную [6].

На данном этапе вопрос разметки частей речи считается не полностью завершенным; все еще существует несколько спорных моментов из-за лингвистических контрастов между английским и кыргызским языками. Если быть более точным, это обычно вызвано наличием или отсутствием определенных синтаксических категорий кыргызского языка, которые не могут быть найдены в структуре английского языка. Данный корпус ориентирован на создание национального корпуса кыргызского языка, а также стать отправной точкой для создания англо-кыргызского/ кыргызско-английского параллельного корпуса [7].

Рисунок 1. Примерные результаты поиска слова «кыргыз» в корпусе кыргызского языка; https://corpora.clarin-d.uni-saarland.de/cqpweb/kyrgyz_20190418/

Синтаксическая разметка кыргызского языка на основе Universal Dependencies (UD). Синтаксическая аннотация - это практика добавления интерпретирующей лингвистической информации в корпусе. Например, одним из распространенных типов аннотаций является добавление тегов или меток, указывающих класс и категории слов, к которому принадлежат (морфологическая аннотация). Это разметка частей речи (POS-tagging), и она может быть полезна, например, для различения слов, которые имеют одинаковое написание, но разные значения или произношение (см. Рисунок 2). Синтаксическая разметка предложений корпуса осуществляется в рамках грамматики зависимостей: синтаксической структурой является ориентированное дерево, узлами которого являются

слова, а каждое ребро направлено от «слова-хозяина» к «слову-слуге» и соответствует некоторому синтаксическому отношению [8].

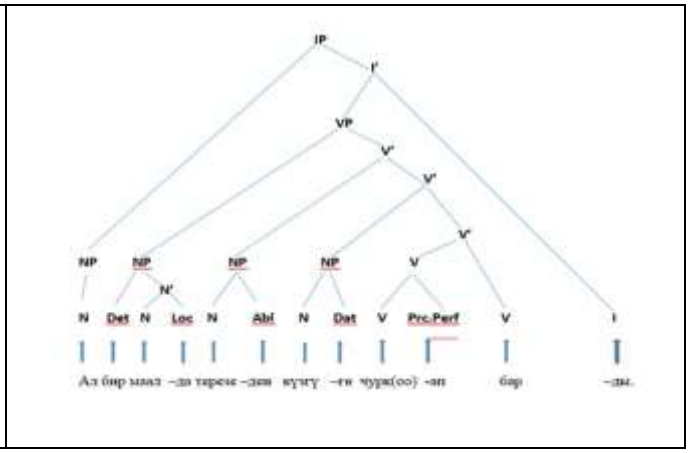
Ал - <prn>, <pers> **бир** – <num>, <card> **маал/да** – <n>, <nn>, <abst>, <sg>, <loc> **терезе/ден** – <n>, <nn>, <concr>, <sg>, <abl> **күзгү/гө** - <n>, <nn>, <concr>, <sg>, <dat> **чурка/п** – <v>, <iv>, <gna_perf> **бар/ды** – <v>, <iv>, <p3>, <ifi>	

Рисунок 2. Образец морфологической и простой синтактической разметок в предложении

На данном этапе совершенствования корпуса кыргызского языка и для изучения структуры предложений и их разметок, мы начали использовать Универсальные Зависимости (UD). Universal Dependencies (UD**)** - это платформа для последовательного аннотирования грамматики (частей речи, морфологических признаков и синтаксических зависимостей) на разных языках [9]. Это проект открытого сообщества, в котором более 300 участников создают почти 200 древостоев на более чем 100 языках. Также, это веб проект, который разрабатывает межъязыковые согласованные аннотации древовидного банка для многих языков с целью облегчения разработки многоязычного синтаксического анализатора, межъязыкового обучения и исследования синтаксического анализа с точки зрения языковой типологии. Схема аннотаций основана на эволюции (универсальных) зависимостей Стэнфорда, универсальных тегов частей речи Google и Interset interlingua для наборов морфосинтаксических тегов. Общая философия заключается в предоставлении универсального списка категорий и руководящих принципов для облегчения последовательного аннотирования схожих конструкций на разных языках, позволяя при необходимости расширения для конкретного языка [10].

На сегодняшний день на официальном сайте UD не имеется примеров и аннотаций на кыргызском языке, что и послужило толчком для новых начинаний в развитии нашего корпуса. На сайте очень мало примеров синтаксических аннотаций тюркских языков, есть разметки на турецком, на татарском и всего 1078 предложений на казахском языке. К сожалению, нет ни единого примера на кыргызском и это показывает уровень вовлеченности лингвистов в мировые новшества. На факультете синхронного перевода КТУ «Манас» сами преподаватели и студенты взялись за построение и отправку

примеров синтаксических разметок на основе УЗ.

Синтаксическая связь между двумя словами является ядром схемы синтаксических аннотаций УЗ. Каждое слово и каждый знак препинания представляют собой пронумерованный узел, начинающийся с 1. Каждый узел зависит от другого узла, который является его управляющим, а узел 1 зависит от КОРНЯ узла (пронумерованного 0), и этот узел является головным. Синтаксические отношения описывают, как зависимые слова зависят от управляющего слова [11].

Предлагаем ознакомиться с синтаксическими разметками примеров на кыргызском и английском языках, выполненных на сайте аннотации UD Annotatrix (https://jonorthwash.github.io/ud-annotatrix/server/public/html/annotatrix.html#1) [12]. На рисунке 3 представлена разметка известной пословицы: «*Койду көсөм баштайт, сөзду чечен баштайт*». Структурно данная пословица является сложно-сочиненным предложением, которое состоит из двух независимых друг от друга предложений, каждый из которых в свою очередь, имеет идентичный порядок слов (Parataxis). Так как эти предложения были объеденены в одно сложное, корнем (root) считается последний предикат/глагол – *баштайт*. И первый предикат, относясь ко второму, выявляет паратаксис (parataxis - «выстраивание рядом»). Паратаксис – это способ выстраивания одинаковых предложений. Также, паратаксис - это литературный прием, в письменной или устной речи, который предпочитает короткие, простые предложения, без союзов или с использованием координирующих, но не подчиняющих союзов [11].

Рассмотрим следующую структуру зависимостей. Связь между словами в УЗ отмечается специальными знаками [10]. Если одно слово зависит от второго, то это будет 1>2, где знак > указывает на зависимость слов. Здесь, в первой части предложения, мы имеем одно простое предложение «*Койду көсөм баштайт*», где «көсөм баштайт» – это главные члены предложения (Noun and Subject), и «койду» – это дополнение (Object), где вторая часть предложения полностью «зеркалит» первую. Синтаксическая разметка данной пословицы была сделана на официальном сайте для создания древовидных зависимостей слов в предложении [12]:

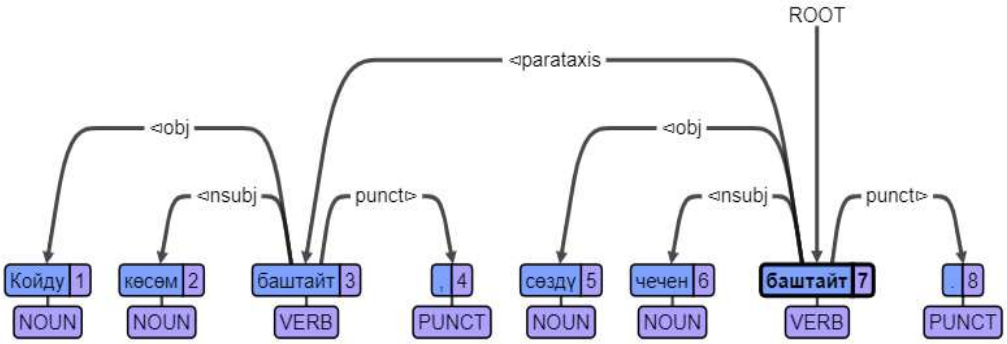


Рисунок 3. Пример разметки пословицы на кыргызском языке

Рассмотрим разметку этой же пословицы в переводе на английский язык. Как видно на рисунке 4, перевод частично совпадает со структурой разметки на кыргызском. Есть незначительные различия, а именно, наличие в английском языке артиклей, которые зависят от существительных, а также характерных структур SVO в английском языке и SOV в кыргызском языке. Однако, стоит отметить, что в литературном и художественном языке эта структура может переставляться для достижения стилистических окрасок, как в пословице «*Койду көсөм баштайт, сөздү чечен баштайт*», где структура предложений – OSV (дополнение, подлежащее и сказуемое).

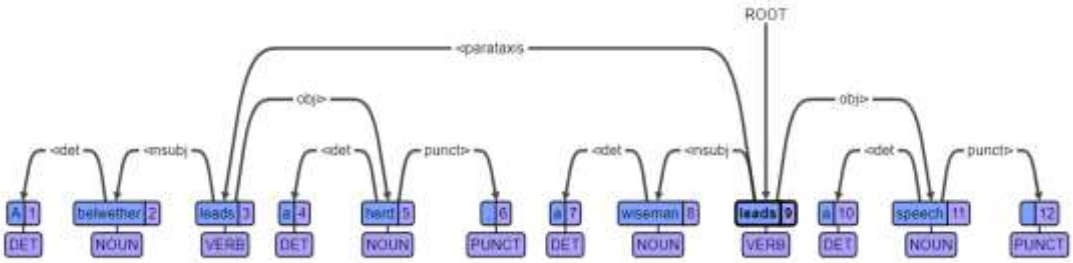


Рисунок 4. Пример разметки пословицы на английском языке

Заключение. В результате исследования выявлены морфологические и синтаксические особенности при адаптации кыргызских грамматических категорий к универсальным символам специальных веб-инструментариев. Произведена разметка синтаксических зависимостей кыргызского предложения на базе УЗ. Введен ряд терминов, связанных с компьютерной обработкой языка и ознакомление с ново созданным кыргызским корпусом. Однако, работа над корпусом считается еще незавершенной; предстоит проделать немало работ для его совершенствования и создания новых параллельных корпусов с кыргызским языком. Надеемся, что это исследование будет способствовать популяризации современного кыргызского языка и использования корпуса кыргызского языка для

различных научных и образовательных целей: изучения кыргызского языка как неродного, голосового распознавания кыргызского языка в гаджетах, интернет-облаках и прочее. Эти небольшие шаги по улучшению корпуса будут способствовать реализации и принятию требований компьютеризации и цифровизации кыргызского языка.

Литература:

1. G. de Walter , «Corpus Linguistics and Linguistic Theory», 2020.

2. McEnery T. и Hardie A., «What is Corpus Linguistics?,» Cambridge Core, 2012. [В Интернете]. Available: https://www.cambridge.org.

3. «Национальный корпус русского языка,» 2003-2022. [В Интернете]. Available: https://ruscorpora.ru/new/.

4. «Корпус кыргызского языка,» Saarland University, the Department of Language and Science , 2019. [В Интернете]. Available: https://corpora.clarin-d.uni-saarland.de/cqpweb/.

5. Kasieva A., Knappen J., Fischer S. и Teich E., «A new Kyrgyz Corpus: sampling, compilation, annotation,» Kyrgyz-Turkish Manas University, Bishkek, Kyrgyzstan, Universität des Saarlandes, Saarbrücken, 04 March 2020. [В Интернете]. Available: https://www.zfs.uni-hamburg.de/dgfs2020/programm/abstracts/dgfs2020-clp-kasieva.pdf.

6. Kasieva A. и Satybekova A., «ЧАСТЕРЕЧНЫЕ РАЗМЕТКИ ДЛЯ НОВОГО КОРПУСА КЫРГЫЗСКОГО ЯЗЫКА», *Вестник КРСУ,* 2020.

7. Kasieva A. и Kadyrbekova A., «Инструментарии для аннотации лингвистического корпуса: Корпус кыргызского языка (Инструментарии Turkic Lexicon Apertium и Penn Treebank),» *Kyrgyz Turkish University "Manas"*, № 20, Saint-Petersburg-Bishkek 2021, 2021.

8. Sekhar Dash N., «Language Corpora Annotation and Processing,» 2021.

9. «Universal Dependencies,» 2014-2021. [В Интернете]. Available: https://universaldependencies.org/.

10. Marneffe et al. , Universal Dependencies, Columbus: Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International, 2021.

11. Thompson A., *Syntactic Parallelism and Structure in Kyrgyz Proverbs,* Pennsylvania, 2021.

1. J. North Washington, «UD Annotatrix,» 2021. [В Интернете]. Available: https://jonorthwash.github.io/ud-annotatrix/server/public/html/annotatrix.html#1.