УДК 811.512.154:004.738.1
DOI: 10.36979/1694-500X-2023-23-6-71-75

АДАПТАЦИЯ ТЕРМИНОВ ВЕБ-ПРОЕКТА УНИВЕРСАЛЬНЫЕ ЗАВИСИМОСТИ НА КЫРГЫЗСКИЙ ЯЗЫК

Г.К. Джумалиева, А.А. Касиева, С.Ж. Мусажанова

Аннотация. Рассматриваются синтаксическая аннотация кыргызского языка и связанные с ней термины посредством языковых корпусов и веб-проектов. При компьютерной обработке языковых корпусов токены, леммы, морфологические теги (POS-теги), синтаксические аннотации слов в текстах сначала помечаются вручную. Далее они автоматически выполняются через компьютерную программу и начинается процесс обработки естественного языка – Natural Language Processing. Однако кыргызские эквиваленты и переводы терминов этой новой отрасли ещё недостаточно освоены. Поэтому одной из целей данной статьи является адаптация терминов веб-проекта универсальные зависимости (Universal Dependencies, далее УЗ) на кыргызский язык. Веб-проект УЗ направлен на выполнение синтаксической аннотации при обработке естественного языка, интерпретации грамматических признаков языков мира и адаптации их к унифицированным наборам языковых признаков, приведённым в единый стандарт. К сожалению, в настоящее время кыргызский язык не включён в список языков, на которых работает веб-проект универсальные зависимости. В связи с этим в данной работе представлены примеры синтаксических аннотаций кыргызского языка с учётом принципов вышеуказанного веб-проекта.

Ключевые слова: языковой корпус; кыргызский корпус; обработка естественного языка; универсальные зависимости; древовидные структуры (treebank); синтаксическая аннотация; морфологические теги.

УНИВЕРСАЛДУУ КӨЗ КАРАНДЫЛЫК ВЕБ-ДОЛБООРУНУН ТЕРМИНДЕРИН КЫРГЫЗ ТИЛИНЕ ЫҢГАЙЛАШТЫРУУ

Г.К. Джумалиева, А.А. Касиева, С.Ж. Мусажанова

Аннотация. Макалада кыргыз тилинин синтаксистик аннотациясы жана ага байланыштуу терминдер тил корпустары жана веб-долбоорлор аркылуу каралат. Тил корпустарын компьютердик иштетүүдө токендер, леммалар, морфологиялык тегдер (POS-тегдер), тексттердеги сөздөрдүн синтаксистик аннотациялары алгач кол менен белгиленет. Андан кийин алар автоматтык түрдө компьютердик программа аркылуу аткарылат жана табигый тилди – Natural Language Processing иштетүү процесси башталат. Бирок бул жаңы тармактын терминдеринин кыргыз эквиваленттери жана котормолору азырынча жетиштүү деңгээлде өздөштүрүлө элек. Ошондуктан бул макаланын максаттарынын бири универсалдуу көз карандылык (Universal Dependencies) веб-долбоорунун терминдерин кыргыз тилине ылайыкташтыруу болуп саналат. Универсалдуу көз карандылык веб-долбоору табигый тилди иштетүүдө синтаксистик аннотацияны аткарууга, дүйнө тилдеринин грамматикалык белгилерин чечмелөөгө жана аларды бирдиктүү стандартка келтирген тилдик белгилердин бирдейлештирилген топтомуна ылайыкташтырууга багытталган. Эгерде бирдейлештирүү мүмкүн болбосо, анда бул тилдин өзгөчөлүгүн сактап калуу керек. Тилекке каршы, учурда кыргыз тили универсалдуу көз карандылык веб-долбоору иштеген тилдердин тизмесине киргизилген эмес. Ушуга байланыштуу бул эмгекте жогоруда аталган веб-долбоордун принциптерин эске алуу менен кыргыз тилинин синтаксистик аннотацияларынын мисалдары келтирилген.

Түйүндүү сөздөр: тилдик корпус; кыргыз корпусу; табигый тилди иштетүү; универсалдуу көз карандылык; дарак түрүндөгү түзүмдөр (treebank); синтаксистик аннотация; морфологиялык тегдер.

ADAPTATION OF THE TERMS OF THE WEB PROJECT UNIVERSAL DEPENDENCIES INTO THE KYRGYZ LANGUAGE

G.K. Dzhumalieva, A.A. Kasieva, S.Zh. Musazhanova

Abstract. This article is aimed at discussing the syntactic annotation of the Kyrgyz language and related terms through language corpora and web projects. During computer processing of language corpora, actions such as tokens, lemmas, morphological tags (POS tags), and syntactic annotation of words in the texts inside them are first marked manually. Then they are automatically executed through a computer (natural language processing - NLP). However, the Kyrgyz equivalents and translations of the terms of this new sphere have not yet been sufficiently mastered. Therefore, one of the goals of this article is to adapt the terms of the Universal Dependencies web project into the Kyrgyz language. The Universal Dependencies web project is aimed at performing syntactic annotation in natural language processing, interpreting the grammar of the world's languages, and adapting them to categories given in accordance with a single standard. Unfortunately, currently, the Kyrgyz language is not included in the list of languages in which the Universal Dependency web project works. In this regard, in this paper, we will discuss the examples that we have prepared as a proposal of syntactic models of the syntactic annotation of the Kyrgyz language for the above-mentioned web project.

Keywords: Language corpus; Kyrgyz corpus; Natural Language Processing; Universal Dependencies; tree structures (treebank); syntactic annotation; morphological tags.

Введение. В эпоху цифровых приложений, гаджетов и с развитием технологий лингвистика не осталась в стороне и появились всевозможные платформы, парсеры и корпуса для осуществления морфологических и синтаксических анализов при обработке естественного языка. Термин *«корпус»* происходит от латинского слова «*corpus*», что означает *«тело»*. Сегодня под словом *«корпус»* мы имеем в виду особый набор текстов определённого языка, диалекта или другого подмножества языка, который используется для лингвистического анализа [1]. Более точное определение состоит в том, что корпус относится в широком смысле к корпусу машиночитаемого текста или набору текстов, выбранных для представления того или иного языка [2]. Существует множество различных типов корпусов. Например, в зависимости от количества представленных языков корпус может быть одноязычным (на одном языке) или многоязычным (на двух или более языках).

Корпус любого языка строится из электронных машиночитаемых текстов, которые в свою очередь содержат токены. **Токен** – это наименьшая единица, из которой состоит корпус, т. е. словоформа/слово, пунктуация, цифра, аббревиатура и всё остальное, что находится между пробелами. А сам процесс разделения каждого токена друг от друга называется **токенизацией** [3].

При создании корпусов язык должен пройти тщательную обработку естественного языка, чтобы компьютер мог распознавать и читать данный язык. **Обработка естественного языка** (**NLP**) – это способность компьютерной программы понимать человеческий язык в устной и письменной форме, называемый естественным языком. Она является одним из компонентов искусственного интеллекта (ИИ). NLP существует уже более 50 лет и уходит корнями в область лингвистики. Это многоэтапная разработка языка, при которой осуществляются различные сложные процессы, как токенизация, лемматизация (процесс приведения словоформы к лемме – словарной форме), парсинг и синтаксическая аннотация предложений [4].

Синтаксический анализ (или **аннотация**) – это автоматический разбор синтаксической структуры естественного языка, синтаксических отношений (в грамматике зависимостей) и частеречной разметки. Синтаксический анализ – одна из ключевых задач компьютерной лингвистики и обработки естественного языка и является предметом исследований с середины XX века с появлением компьютеров [5]. Для выполнения вышеперечисленных задач в лингвистике употребляется комплекс программных модулей **(парсеры)**, который выполняет разбор линейной последовательности лексем языка [6], а, собственно, сам грамматический процесс, который предполагает разбиение текста на составные части речи (Parts of speech tagging) с объяснением формы, функции и синтаксических отношений каждой части, называется **парсингом** (происходит от латинского

«*pars*», означающего *«часть речи»*). В мире существуют десятки разных парсеров для синтаксического разбора предложений при обработке естественного языка [7]. Одним из таких парсеров является **Universal Dependencies** (URL: https://universaldependencies.org/). Онлайн-платформа, известная как **универсальные зависимости**, имеет открытый исходный код и включает в себя последовательные аннотации кросс-лингвистических древовидных структур (treebanks). Это уникальная программа, разработанная американскими учёными в 2004 году для Стэнфордского университета на основе межъязыковых морфосинтаксических тегов, универсальных POS-тегов Google и универсальных зависимостей Стэнфорда [8]. В УЗ для синтаксического анализа употребляется специальная онлайн-платформа UD Annotatrix, в которой можно выполнять древовидные структуры предложений на всех языках мира [9].

Древовидная структура (**treebank**) – это особый подход к анализу предложений, применяющий схемы и показывающий связи фраз в предложении. Термин «treebank» был введён лингвистом Джеффри Личем в 1980-х годах. Древовидные структуры часто создаются на основе корпуса, который уже аннотирован морфологическими тегами частей речи. В свою очередь, эти структуры иногда дополняются семантической или другой лингвистической информацией. **Treebanks** могут быть созданы полностью вручную, когда лингвисты аннотируют каждое предложение с синтаксической структурой, или полуавтоматически, когда синтаксический парсер назначает некоторую синтаксическую структуру, которую лингвисты затем проверяют и при необходимости корректируют. Таким образом, универсальная зависимость (УЗ) работает полуавтоматически, создавая древовидные структуры предложений, выполняемые посредством соблюдения основных принципов, где словам присваивают морфологические признаки, далее эти слова вступают в синтаксические отношения, которые отражают аннотации УЗ.

Анализ. Рассмотрим на примере синтаксическую аннотацию предложения на кыргызском языке (рисунок 1).

Данное предложение имеет очень необычную структуру, где отсутствует подлежащее и сказуемое состоит из двух слов (главного и вспомогательного). Зависимость подлежащего и сказуемого в УЗ отмечается **«nsubj»**, которое может возглавляться существительным, или это может быть местоимение или относительное местоимение или, в контексте эллипсиса, другие части речи, как прилагательное. Однако это безличное предложение имеет сложную структуру зависимостей. Унифицированные символы стрелками указывают на то, как одно слово зависит от другого в предложении. **«ROOT»** – это корень предложения, сказуемое, так называемое самое значимое слово, которое несёт собой смысл всего предложения. Данная грамматическая связь указывает на корень предложения, которым является слово «таанышууга». Узел ROOT индексируется цифрой 0, так как нумерация реальных слов в предложении начинается с 1 и затем каждый токен отмечается номерами 2, 3, 4 и т. д. Без данного компонента – корня предложения – высказывание не имело бы никакого смыслового значения. **«Aux»** – это вспомогательный глагол, где его роль играет слово

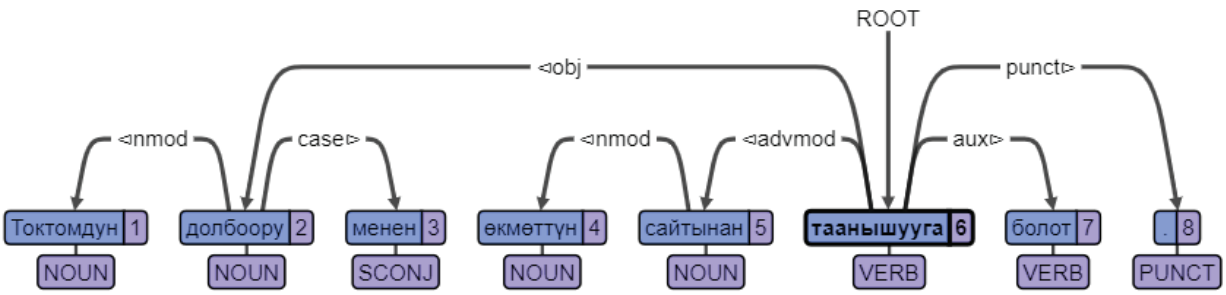

Рисунок 1 – Пример синтаксической аннотации предложения из кыргызского корпуса

«болот». «Aux» модифицирует глагольный предикат, добавляя информацию, относящуюся ко времени, к роду, лицу, числу или падежу. *«Токтомдун долбоору»* и *«өкмөттүн сайтынан»* между собой связываются зависимостью «nmod», которое является номинативной фразой или номинативом, модифицирующим/определяющим главное слово другого номинатива (существительного) маркером особого склонения или без него. **«Conj»** – это отношение между двумя элементами, связанными координацией или предлогом, такой как «and», «or», и т. д. В случае кыргызского языка все существительные объединяются с помощью послелога *«менен»*, который может выполнять три функции: instrumental (творительный падеж), sociative (социативный падеж со значением совместности), conjunctive (соединительный падеж). **«Case»** – падеж, обычно является флективным признаком существительных и в зависимости от языка других частей речи (местоимений, прилагательных, числительных, глаголов), которые обозначают согласование с существительными. Однако падеж в кыргызском языке указывается не предлогами, а к существительным просто добавляются падежные окончания, которые могут выражаться послелогами. Связь «case» в этом предложении играет зависимость *«долбоору менен»*, которая указывает на роль дополнения. Падеж помогает определить роль существительного в предложении, особенно в языках со свободным порядком слов (агглютинативные языки, как кыргызский). Например, в языках с фиксированным порядком слов эти функции различаются просто позициями существительных в предложении [10].

«Obj» – это второй по значимости аргумент глагола после субъекта. Обычно это существительное, обозначающее объект, на который направлено действие или который претерпевает изменение состояния или движения. В случае данного предложения *«долбоору менен таанышууга болот»* зависимость указывает на объект, над которым происходит действие.

В стандартном списке символов УЗ **«punct»** – это знаки препинания; неалфавитные символы и группы символов, используемые во многих языках для разграничения языковых единиц в печатном тексте. В УЗ «punct» могут быть все символы, такие как точка, запятая, вопросительный и восклицательный знаки, двоеточие, и др. Токены, обозначаемые «punct», всегда присоединяются к смысловым словам, корню предложения «ROOT» (за исключением случаев эллипсиса) и никогда не могут иметь зависимых слов.

Выводы. Таким образом, проект универсальных зависимостей позволяет унифицировать морфосинтаксическую схему анализа языков мира (см. таблицу 1). Он позволяет упростить межъязыковые исследования, унифицирует межъязыковые лингвистические типологии, обеспечивает основу для построения автоматизированных многоязычных систем и предоставляет универсальный анализатор текста [10].

Заключение. В результате изучения разных источников раскрыты описания синтаксической аннотации предложения, парсинга и УЗ, задачей

Таблица 1 – Глоссарий адаптации терминов

Английский	Русский	Кыргызский
corpus	корпус	корпус
token	токен	токен
tokenization	токенизация	токенизация
natural language processing	обработка естественного языка	табигый тилди иштетүү
syntactic annotation	синтаксическая аннотация	синтаксистик энтектөө
parser	парсер	парсер
parsing	парсинг	парсинг
universal dependencies	универсальные зависимости	универсалдуу багыныңкылык
treebank	древовидная структура/дерево зависимостей/банк структур	дарак түрүндөгү структура

которого является унификация синтаксических грамматик языков мира. Продемонстрирована синтаксическая аннотация предложения на кыргызском языке на примере древовидной структуры (treebank) и даны примерные варианты терминов, используемых в рамках УЗ. И именно через данную платформу УЗ работа по углубленному изучению кыргызского языка и участию в этом мировом проекте позволит вывести изучение грамматики кыргызского языка на новый уровень.

Благодарность. *Выражаем огромную благодарность профессору Джонатан Норт Вашингтону из колледжа Свартмоор, шт. Пенсильвания за поддержку и предоставление веб-платформы UD Annotatrix: URL: https://jonorthwash.github.io/udannotatrix/src/server/public/html/annotatrix.html#1*

Поступила: 28.12.22; рецензирована: 12.01.23; принята: 16.01.23.

Литература

1. *Грудева Е.В.* Корпусная лингвистика: учеб. пособие. 2-е изд., стер. / Е.В. Грудева. М.: Флинта, 2012. 165 с.
2. *McEnery T.* Corpus Linguistics: An Introduction / T. McEnery, A. Wilson // 2nd Edition, Edinburgh University Press. Edinburgh. 2001. 122 p. URL: https://uogbooks.files.wordpress.com/2014/10/tony_mcenery_andrew_wilson_corpus_linguisticsbook4you-org.pdf (дата обращения: 25.11.2022).
3. *Hagiwara M.* Real-World Natural Language Processing / M. Hagiwara // Manning Publications, 2021. URL: https://vk.com/doc255577237_622351969 (дата обращения: 25.11.2022).
4. *Lutkevich B.* What is natural language processing? An introduction to NLP / B. Lutkevich, E. Burns // Techtarget.com, 02 Mar-2021. URL: https://searchenterpriseai.techtarget.com/definition/natural-languageprocessing-NLP (дата обращения: 22.11.2022).
5. Ruscorpora / Национальный корпус русского языка. URL: https://ruscorpora.ru (дата обращения: 14.10.2022).
6. *Гаршина В.В.* Разработка лингвистического парсера русского языка / В.В. Гаршина, Ю.А. Богоявленская // Вестник ВГУ. Системный анализ и информационные технологии. 2012. № 2. С. 174–182. URL: http://www.vestnik.vsu.ru/pdf/analiz/2012/02/2012-02-29.pdf (дата обращения: 14.10.2022).
7. *Jurafsky D.* Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition / D. Jurafsky, J. Martin // 3rd Edition draft, 2020. URL: https://web.stanford.edu/~jurafsky/slp3/ (дата обращения: 26.11.2022).
8. *Nivre J.* Universal Dependencies / J. Nivre, D. Zeman, F. Ginter, and F. Tyers // In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts, Valencia, Spain. Association for Computational Linguistics, 2017. URL: https://aclanthology.org/E17-5001.pdf (дата обращения: 05.10.2022)
9. *Washington J.N.* Annotatrix. URL: https://jonorthwash.github.io/ud-annotatrix/src/server/public/html/annotatrix.html#1.
10. *Люкина Е.В.* Использование универсальных зависимостей при грамматическом разборе многоязычного текста (на примере безличного предикатива) / Е.В. Люкина // Вестник НГУ. Сер.: Лингвистика и межкультурная коммуникация. Новосибирск, 2018. Т. 16. № 2. С. 19–33. URL: https://cyberleninka.ru/article/n/ispolzovanie-universalnyh-zavisimostey-pri-grammaticheskom-razbore-mnogoyazychnogo-teksta-na-primere-bezlichnogo-predikativa/viewer (дата обращения: 14.10.2022).