КОРПУСНАЯ ЛИНГВИСТИКА И КОРПУСНЫЕ ИССЛЕДОВАНИЯ

УДК 81’322.2

ISSUES OF KYRGYZ SYNTACTIC ANNOTATION WITHIN THE UNIVERSAL DEPENDENCIES FRAMEWORK

Aida Kasieva[1], Gulnura Dzhumalieva[1], Anna Thompson[2], Murat Jumashev[3], Bermet Chontaeva[4], Jonathan Washington[5]
[1]Kyrgyz_Turkish Manas University, Bishkek, Kyrgyzstan,
[2]Independent researcher, Leeds, UK
[3]Independent researcher, Bishkek, Kyrgyzstan
[4]Universität Tübingen, Tübingen, Germany,
[5]Swarthmore College, Swarthmore, USA,
aida.kasieva@manas.edu.kg, gulnur.jumalieva@manas.edu.kg,
thompsonannad@gmail.com, jumasheff@gmail.com,
bermet.chontaeva@student.uni-tuebingen.de,
jonathan.washington@swarthmore.edu

This paper examines key issues encountered in syntactic annotation work for a forthcoming Universal Dependencies (UD) corpus of Kyrgyz. It presents an overview of the corpus creation process, including sentence sampling from the Manas-UdS Kyrgyz corpus and manual annotation using UD guidelines. The corpus contains over 1600 tokens across 230 sentences sampled from literary and news domains. Four central issues in Kyrgyz UD annotation are then discussed in-depth: copula tokenization, categorization of “small words” like да and керек, null-headed clauses (including relative clauses, and -DAGI and -NIKI constructions), and differentiating inflection vs. derivation. For each issue, multiple analysis options are weighed, including contrasting the approach in prior Turkic UD treebanks. Copula analysis compares subject agreement morphology as dependent subtokens vs independent words. The discourse and intensifier functions of да are examined to determine optimal POS and dependency labels. Strategies for representing implicit nominal heads in relative clauses and genitive constructions are evaluated. Criteria for categorizing productive derivational morphology as inflectional cases vs separate words are outlined. Throughout, examples illustrate annotation decisions and dependency graphs. Comparisons are made to the analysis of related phenomena in existing UD treebanks for Kazakh [Tyers & Washington 2015, Makazhanov et al. 2015], Turkish, and the small Kyrgyz UD corpus [Benli, 2023]. The work identifies ongoing challenges in representing Kyrgyz syntax within UD, while developing an improved annotated resource. It highlights issues where UD guidelines exhibit limitations for Turkic languages, providing analysis to advance understanding of best practices for Kyrgyz and related languages.

Keywords: Kyrgyz, syntax, annotation, Universal Dependencies.

ПРОБЛЕМЫ КЫРГЫЗСКОЙ СИНТАКСИЧЕСКОЙ АННОТАЦИИ В ФРЕЙМВОРКЕ UNIVERSAL DEPENDENCIES

Касиева А.[1], Джумалиева Г.[1], Томпсон А.[2], Юмашев М.[3], Чонтаева Б.[4], Джонатан В.[5]

[1]Кыргызско-Турецкий университет Манас, Бишкек, Кыргызстан,
[2]Независимый исследователь, Лидс, Великобритания
[3]Независимый исследователь, Бишкек, Кыргызстан
[4]Тюбингенский университет, Тюбинген, Германия,
[5]Суортморский колледж, Суортмор, США,
aida.kasieva@manas.edu.kg, gulnur.jumalieva@manas.edu.kg,
thompsonannad@gmail.com, jumasheff@gmail.com,
bermet.chontaeva@student.uni-tuebingen.de,
jonathan.washington@swarthmore.edu

В данной статье рассматриваются основные вопросы, возникающие в процессе работы над синтаксической аннотацией для разрабатываемого корпуса кыргызского языка на базе универсальных зависимостей (УЗ). Представлен обзор процесса создания синтаксического корпуса, предложения для которого отобраны из корпуса кыргызского языка “Manas-UdS”. Синтаксическая аннотация корпуса выполняется вручную в соответствии с рекомендациями УЗ. Данный корпус содержит более 1600 токенов в 230 предложениях, отобранных из текстов художественных произведений и новостей. Подробно рассматриваются четыре основные проблемы аннотирования УЗ кыргызского языка: токенизация копулы (глаголы-связки), категоризация «служебных слов», таких как “да” и “керек”, предложения с отсутствующим главным элементом (null-head), включая относительные предложения, конструкции с -ДАГЫ и -НЫКЫ, а также разграничение между флексией и деривацией. Для каждого случая рассматриваются различные варианты анализа, включая сравнение подходов, применяемых в существующих тюркских УЗ-трибанках. Анализ копулы позволяет сравнить морфологию согласования подлежащего в качестве зависимых субтокенов с независимыми словами. Функции дискурса и усилителя “да” рассматриваются для определения подходящих для него частей речи и соответствующих аннотаций зависимостей. Оцениваются стратегии представления имплицитных номинативных элементов в относительных предложениях и генитивных конструкциях. Изложены критерии, определяющие, следует ли классифицировать продуктивную деривационную морфологию как случай словоизменения или как отдельные слова. Примеры иллюстрируют используемые модели аннотаций и диаграмм зависимостей. Проводятся сравнения анализа подобных случаев в существующих УЗ трибанках для казахского [Tyers & Washington 2015, Makazhanov, 2015], турецкого и небольшого корпуса УЗ для кыргызского языка [Benli, 2023]. В работе выявляются проблемы, связанные с представлением синтаксиса кыргызского языка в рамках УЗ при разработке улучшенного варианта аннотированного ресурса. Наряду с выявлением моментов, в которых руководство по УЗ демонстрирует ограничения в рекомендациях по использованию

УЗ для тюркских языков, также предлагаются более оптимальные варианты анализа для кыргызского и смежных языков.

Ключевые слова: Кыргызский язык, синтаксис, аннотация, универсальные зависимости

I. Introduction

This paper examines issues that have arisen in syntactic annotation work for a forthcoming Universal Dependencies (UD) corpus of Kyrgyz, a Turkic language of Central Asia. We lean on prior research on Kyrgyz syntax and existing UD corpora of Turkic languages as a foundation, and use existing Kyrgyz textual analysis tools and UD annotation tools for our work. Manually annotated syntactic data is an invaluable resource for understanding the grammatical patterns and constructions of a language.

The creation of a UD Kyrgyz treebank will support more in-depth investigation into the syntax and morphology of Kyrgyz within a cross-linguistically consistent framework. In addition, annotated syntactic data is essential for developing accurate natural language processing tools like part-of-speech taggers and parsers. The release of a high-quality, comprehensive UD treebank for Kyrgyz will fill a crucial gap, enabling the training of NLP models for syntactic and morphological analysis, machine translation, information retrieval and more. Currently the resources available for syntactically parsing Kyrgyz text are limited. This work seeks to address this need and provide a valuable annotated dataset that can serve as training and evaluation data for Kyrgyz language technologies.

Our treebank draws sentences from a broader range of domains contained in the Manas-UdS Kyrgyz corpus. Second, we provide an in-depth analysis of major syntactic issues that have arisen during the annotation of copula tokenization, null-headed clauses, and differentiating derivation from inflection. We extensively compare potential solutions for each issue, weighing the probability of their occurrences in Kyrgyz. Third, we contrast our analysis and annotation decisions with those made previously, particularly the existing UD treebanks for Kyrgyz [Benli, 2023]. The creation of a larger treebank and thorough examination of ongoing annotation challenges advances understanding of applying UD conventions to Kyrgyz and builds a higher quality resource to support future parsing and NLP applications.

Section 2 discusses background on UD, its application to Turkic languages, and syntactic research into Kyrgyz. Section 3 overviews the annotation work of the authors to date.

In Section 4, a range of issues encountered in annotation are discussed. These include copula tokenisation, the treatment of difficult-to-categorise «small» words, null-headed clauses (including relative clauses, and *-DAGI* and *-NIKI* constructions), and decisions regarding inflection versus derivation. Section 5 concludes and proposes future work.

II. Background

The Universal Dependencies project aims to develop cross-linguistically consistent treebank annotation for many of the world's languages [Nivre, 2016]. It represents predicate-argument structure through labeled dependency parses, providing common guidelines for annotation across languages. The consistency enables cross-linguistic learning and analysis. The sentences in the US are represented through directed acyclic graphs, with words as nodes and grammatical relations as labeled edges. UD guidelines strive for consistency across languages, while allowing language-specific extensions; the quality and coverage of UD resources varies across languages.

Several UD treebanks have been developed for Turkic languages, including Turkish, Uyghur, and Kazakh, as well as a recent treebank for Kyrgyz.

Tyers and Washington [Tyers, 2015] describe the development of the first free and open-source dependency treebank for Kazakh, which they released using UD v1 annotation standards. At the time of publication, the treebank contained 402 sentences from open-source and public domain texts to ensure free availability and extensibility (it is now larger). The texts were first morphologically analysed and disambiguated using existing resources for Kazakh [Washington et al. 2014], and were then manually annotated for dependency syntax. The authors further discuss several linguistic issues in Kazakh focusing on their analysis in UD, including functions of case morphemes, derivations, non-finite clauses, and copulas. The decisions of annotation are outlined, like marking the copula as a dependent and last conjunct as the head in coordination. Verbal nouns, adjectives, and adverbs are annotated for their functions as subjects, modifiers, or clausal complements. Their preliminary parsing experiments showed 63.9% LAS and 74.7% UAS with structural features, comparable to other small treebanks. This treebank has since been converted to be in line with UD v2.

Makazhanov et al. [Makazhanov, 2015] conducted their study based on 300 sentences randomly selected from the closed source Kazakh Language Corpus [Makhambetov et al. 2013]. Their work on syntactic annotation revealed several challenges. First, they had difficulty categorizing the analytic negation markers *жоқ* and *емес*, ultimately opting to classify them as copulas. Second, their dataset did not contain non-relative (acl:relcl) examples of clausal noun modifiers, resulting in annotations with no specified clausal noun modifier relation. Lastly, they faced challenges in ensuring accurate dependency relations, particularly in complex phrases like *үлкен үйдегілер* 'those in the big house'. In such cases, directly attaching the adjective 'big' to 'those in the house' led to a misrepresentation of the intended meaning, highlighting the need for consistent tokenisation and annotation conventions.

Tyers et al. [Tyers, 2017] present an early assessment of UD guidelines for Turkic languages. They highlight areas of cross-linguistic consistency, and note discrepancies between guidelines for Turkish, Kazakh, and Uyghur. Open issues discussed include tokenization, differentiating core arguments, complex predicates, and copula usage. Our work builds on their assessment, tackling similar issues for our Kyrgyz UD treebank.

Aili et al. [Aili, 2018] took steps to extend Universal Dependencies (UD) resources to Uyghur. They mapped the treebank's labeling scheme to UD labels, making structural changes like marking auxiliaries and copulas as dependents. Some UD relations were introduced for Uyghur-specific syntax like modifier emphasis (advmod:emph). Aili et al. also defined new labels needed to represent complex Uyghur structures concisely within UD, including compound reduplication (compound:redup). Their work demonstrates both adapting UD's universal principles to Uyghur and extending UD conventions as required for the language.

Four treebanks have also been published for Turkish, along with a number of academic papers associated with them [e.g., Sulubacak et al. 2016, Çetinoğlu and Çöltekin 2022].

Sulubacak et al. [Sulubacak, 2016] converted the IMST Treebank (Turkish), originally available in the CoNLL-X data format, to the CoNLL-U format in compliance with UD standards. Utilizing the Inflectional Group (IG) formalism [Oflazer 1999; Hakkani-Tür et al. 2002], the authors segment orthographic tokens into morphosyntactic words at derivational boundaries. They provide comprehensive

mapping rules for converting both morphological features and dependency relations to align with UD standards. The paper also discusses the challenges of annotating non-projective sentences, which led to a slight drop in labeled attachment scores. The authors test their methodology on the UD version of the IMST Treebank, providing valuable metrics on its effectiveness, achieving a labeled attachment score (LAS) of 81.41% and an unlabeled attachment score (UAS) of 85.48%.

Çetinoğlu and Çöltekin [Çetinoğlu, 2022] present the Turkish-German SAGT code-switching treebank. It contains rich linguistic annotations including language IDs, lemmas, POS tags, features, and dependency relations. The SAGT treebank is one of the few publicly available resources for studying code-switching between German and Turkish. Special care was taken during annotation to handle multilingual consistency and informal language. Features like CSID indicate code-switching type (intra-lexical, intrasentential). Data was collected from conversations between 20 Turkish-German bilingual students and annotated using UD monolingual treebanks. The treebank comprises 2,184 sentences and 37,233 tokens after segmentation. Most annotation differences result from divergent grammatical traditions, not linguistic discrepancies. Challenges identified include consistent multilingual annotation and informal language. Proposed solutions involve tailored guidelines, multiple annotation layers, and contextualized annotation. Iteratively identifying and fixing errors is important since code-switching complexity produces more annotation errors than monolingual treebanks. Overall, this paper introduces an invaluable annotated resource to spur advances in code-switching analysis.

A UD corpus has also been released for Tatar [Taguchi 2022].

These existing UD corpora provide a useful starting point, as models. However, many open questions remain regarding UD annotation for Turkic languages.

As an agglutinative Turkic language, Kyrgyz exhibits flexibility in its word order, including both head-initial and head-final structures. Kyrgyz syntax has been the focus of few studies, with some examination of relative clauses [Imanalieva 2015] and other constructions. The Universal Dependencies framework aims to represent syntactic variation across diverse languages. While UD has some bias toward head-initial order, it can also model head-final structures where attested. Capturing the word order variation found in

Kyrgyz thus presents an interesting test case for dependency annotation under UD guidelines. Further research will be valuable for assessing how well UD accommodates the syntactic patterns of Kyrgyz.

Kyrgyz, as a morphologically rich language, pushes the limits of the guidelines for phenomena like non-canonical word order and complex predicate formations [Thompson 2021].

The Kyrgyz language currently has limited syntactic resources available in the UD framework. As of the UD v2.12 release, the recently added Kyrgyz UD treebank [Benli 2023] contains only 781 sentences and its domain is mainly news headlines and stories selected from Kyrgyz novels and news websites. Details of annotation decisions are not discussed in depth.

Dzhumalieva et al. [Dzhumalieva, 2023] investigate the challenges and opportunities of syntactic annotation for the Kyrgyz language within the UD framework. They propose the adaptation of relevant terminology into Kyrgyz and outline their initial steps in manual tagging of tokens, lemmas, and POS-tags, laying the groundwork for future automated Natural Language Processing tasks. A central focus is syntactic analysis and treebank annotation using the Universal Dependencies framework.

Musazhanova et. al [Musazhanova, 2023] discuss an effort in the syntactic annotation of the Kyrgyz language using UD. The paper offers annotation examples of Kyrgyz sentences and reveals that the Kyrgyz language's grammatical categories haven't been fully explored within the UD framework. The syntactic analysis examples provide insight into adapting Universal Dependencies standards for Kyrgyz. It highlights a significant gap in computational linguistics for Kyrgyz and lays the groundwork for future research on annotated corpus development.

Washington et al. [Washington, 2012] present a finite-state morphological transducer for Kyrgyz. At publication, the lexical foundation covered over 8,000 stems across major word classes; it now covers over 15,000 stems. While intended for machine translation, the transducer may also be used to aid morphological analysis for syntactic parsing. Our UD annotation experience suggests that this resource may need extension to handle complex verbs and other phenomena.

While valuable related work exists on Turkic and specifically Kyrgyz UD, many issues persist in developing high-quality UD-annotated resources for Kyrgyz. Our paper aims to advance understanding of these syntactic annotation challenges through analysis

of a larger, more diverse UD Kyrgyz treebank. To this end, we take steps towards creating an improved Kyrgyz UD corpus through manual annotation of new data.

3. Corpus Development

This section describes our corpus creation process, including sentence sampling, and annotation workflow, as well as corpus statistics and metadata. It also presents the current state and future plans for the corpus.

For completion of undergraduate linguistics coursework, Thompson [Thompson, 2021] completed a thesis on the relationship between syntactic structure and syntactic parallelism using 85 randomly selected Kyrgyz proverbs, building on work done previously for a course project in Washington's Structure of Kyrgyz course. The workflow began with the analysis of the proverbs using the Apertium Kyrgyz morphological transducer [Washington, 2012]. The proverbs were then manually annotated using Universal Dependencies guidelines. This allowed for analysis of the corpus of proverbs to engage with the common terms and categories from previous research about syntactic parallelism and proverb structure. The analysis categorized proverb syntax and identified patterns, like the association between parataxis relations and syntactic parallelism. This corpus of 85 proverbs was made publicly available under an open-source license, constituting the first freely available dependency-annotated corpus of Kyrgyz, and constitutes part of our corpus as well.

Building on this work, Kasieva and Dzhumalieva, along with their students in the Translation department of Kyrgyz-Turkish Manas University, extracted and manually annotated sentences from the 2M-word Manas-UdS Kyrgyz corpus [Kasieva et al. 2020], which was compiled from Kyrgyz literary works and the state newspaper "Erkin-Too."[1], The literary portion was drawn from 12 short stories and 3 novels, selecting sentences with a range of syntactic constructions. The news portion sampled sentences from 15 articles spanning different topics. Care was taken to extract sentences covering diverse lexical content and syntactic phenomena.

New Kyrgyz sentences were manually annotated using the UD Annotatrix interface [Tyers et al. 2018]. Annotations were completed by various combinations of the authors of this paper, often building

[1] https://erkin-too.kg/

on student work. The open-source tool UD Annotatrix provides an interface designed for UD annotation and validation. Annotators followed the UD guidelines, referring to prior analyses of Kyrgyz structures [Thompson, 2021]. Disagreements were discussed and resolved to reach consensus. Inter-annotator agreement was over 90% by the end of the process.

Designed for manual annotation of Universal Dependencies (UD), UD Annotatrix was a valuable asset for creating this Kyrgyz syntactic corpus. The tool handles annotation guidelines such as two-level segmentation schemes, and provides validation feedback. It allows for customization of guidelines specific to the Kyrgyz language and lists and auto-completes language features (e.g., POS and dependency relations). Linguists working with Kyrgyz can utilise its built-in features, including automated parsing and dependency visualization, for a streamlined annotation process. This ensures the creation of a comprehensive and consistent Kyrgyz syntactic corpus.

The resultant corpus contains 2456 tokens across 332 sentences. Sentence lengths range from 5 to 35 tokens, with an average of 14 tokens. The vocabulary size is 829 unique words. Morphological features and universal POS tags were applied using the morphological transducer developed by Washington et al. [Washington, 2012].

Work is underway to finalize annotation, add detailed metadata, expand the corpus size, and prepare submission to the Universal Dependencies project.[1] In future work, we hope to increase the domain diversity by sampling scientific articles, spoken dialogues, and social media text. This will provide a robust annotated corpus to support NLP research on the Kyrgyz language.

4. Issues of interest

This section identifies challenges in applying UD to Kyrgyz by presenting recurring issues that have arisen during annotation. Not all challenges encountered are discussed in this paper.

In our data, we encountered challenges like copula tokenisation (4.1), the treatment of difficult-to-categorise «small» words (4.2), and null-headed clauses (including relative clauses, and *-DAGI* and *-NIKI* constructions) (4.3). We also had to make decisions regarding inflection versus derivation (4.4).

[1] The corpus is currently available at https://github.com/apertium/apertium-kir/tree/main/corpora.

4.1. Copula tokenisation

In Kyrgyz there are several strategies used to form copula sentences.

In non-past-tense copula sentences, the normal strategy is to add a subject agreement morpheme to the predicate, as in (1).

(1) *Мен сенин үйүңдөмүн.*
men senin üy-(I)ŋ-DA-MIn.
I your house-POSS.2SG-LOC-COP.NPST.1SG
'I'm at your house.'

In the past direct tense, an apparently irregular form of a defective verb э- is used, as in (2).

(2) *Мен сенин үйүңдө элем.*
men senin üy-(I)ŋ-DA ele-m.
I your house-POSS.2SG-LOC COP.PST.DIR-1SG
'I was at your house.'

There is additionally an irregular-looking past verbal noun form of э-: экен (cf. expected *эген), as shown in (3).

(3) *Мен сенин үйүңдө экенимди билиптирсиң.*
men senin üy-(I)ŋ-DA eken-(I)m-NI bil-(I)ptIr-sIŋ.
I your house-POSS.2SG-LOC COP.VN-POSS.1SG-ACC know-PST.IDF-2SG
'You knew that I was at your house.'

No other forms of this defective verb exist; missing forms include various tenses and non-finite forms, as well as negative forms.

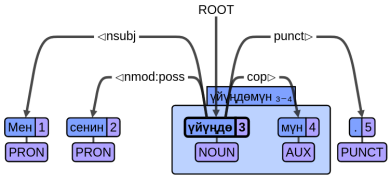
We treat the non-past copula subject agreement morphemes as if they were cliticised forms of the defective copula verb, with lemma э and POS tag AUX, and as a subtoken of the space-delimited «word» that they are part of. Despite the fact that they have an unrelated etymology from the defective copula verb, there are several reasons we believe this approach is advantageous:

1. The defective copula verb does not have non-past forms. This approach allows the non-past agreement morphemes to fill that gap in the paradigm.

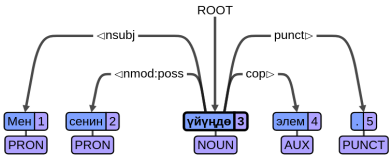
2. This approach allows non-past and direct past to have similar analyses, as shown in Graphs 1 and 2.

3. This approach prevents the problem of having multiple person/number/formality marking on a singular noun, as otherwise would be necessitated in (1).

4. This approach allows the morphemes to be labelled as copula.



Graph 1. UD graph of sentence (1) depicting a non-past copula construction



Graph 2. UD graph of sentence (2) depicting a direct past copula construction

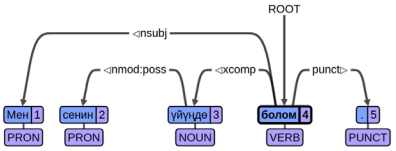
For the sake of consistency, we have chosen to analyse a separate copula subtoken even in the third person (singular and plural), where it has no orthographic content. It would be possible to leave this subtoken out of the annotation, as Tyers and Washington [Tyers, 2015] did for Kazakh, but it would have the disadvantage of then having no indication of subject agreement.

In parts of the paradigm of э- where forms are non-existent, the verb бол- 'be, become' is used instead. In fact, the verb бол- can be used in certain contexts in place of forms of э-. For example, sentences (4) and (5) can have the same meanings as (1) and (2), respectively.

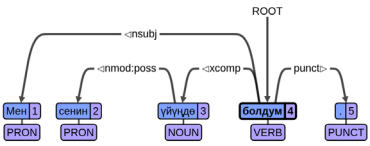
(4) *Мен* *сенинүйүңдө* *болом.*
men senin üy-(I)ŋ-DA bol-E-m.
I your house-POSS.2SG-LOC be-NPST-1SG
'I am (/will be) at your house.'

(5) *Мен* *сенин* *үйүңдө* *болдум.*
men senin üy-(I)ŋ-DA bol-DI-m.
I your house-POSS.2SG-LOC be-PST.DIR-1SG
'I was at your house.'

While these are essentially copula constructions, *бол-* is a regular non-defective lexical verb, and so we treat it as such. In annotation of these sentences, then, *бол-* is annotated as a VERB, and the predicate as an xcomp dependent of it, as shown in Graphs 3 and 4.



Graph 3. UD graph of sentence (4) depicting a non-past *бол-* verbal construction



Graph 4. UD graph of sentence (2) depicting a direct past *бол-* verbal construction

This has the disadvantage of semantically and morphologically very similar structures being treated as having different syntax. Tyers and Washington [Tyers, 2015] opt instead to treat *бол-* constructions the same as *э-* copula constructions.

Benli [Benli, 2023] annotates the following types of copula constructions in the following ways:

• When subject-agreement morphemes occur on non-verbal predicates, the noun or adjective comprising the final word in the predicate is analysed as having person features of the subject, and is sometimes misanalysed as a verb.

• Forms of *эле* are given the POS tag VERB and are treated as compound:svc (elements of serial-verb constructions) dependents of the non-verbal predicate.

• Complements of *бол-* are analysed as amod dependents.

It is not clear what the reasoning for these analyses might be.

4.2. «Small» words

This section addresses the analysis of several «small» words that originally presented difficulties for annotation: *да* (§4.2.1), *эле* (§4.2.2), *бар* and *жок* (§4.2.3), and *керек* (§4.2.4).

4.2.1. *да*

In Kyrgyz, there are several distinguishable uses of the word *да*, likely constituting several distinct lexical words. These are the uses, as delimited by the authors:

1. Post-predicate «modal particle». In this use, *да* indicates that the speaker(s) is making a statement whose truth value they believe to be evident to the interlocutor(s), but which needs to be asserted to explain something else. An example of this from the corpus is given in (6).

(6) *Натыйжалар жарыяланыптыр да, ээ?*
'The results have been announced, haven't they?'

2. Conditional intensifier. In this use, *да* adds intensity to a conditional adverbial clause, translating to English roughly as «even» in uses as «even if». An example of this from the corpus is given in (7).

(7) *Оозу кыйшык болсо да, байдын уулу сүйлөсүн.*
'Even if his mouth is crooked, let the rich person's son speak.'

3. General contrastive intensifier. In this use, *да* adds a contrastive focus to the preceding element, which may constitute a wide range of phrase types. It can be translated to English as «even». An example of this from the corpus is given in (8).

(8) *Тамашада да чындыктын үлүшү бар.*
'Even in a joke is some element of truth.'

4. General conjoining adverb. In this use, *да* adds the sense that what is being said about the preceding phrase is true in addition that same situation regarding a parallel phrase. It can be translated to English as «also» or «too». This meaning and the preceding one may often both be interpreted in a single example. An example of this from outside the corpus is given in (9).

(9) *Атам да каршы болду.*
'My father was also against it.' (or: 'Even my father was against it.')

5. Correlative conjunction. In this use, *да* is used twice, with two parallel phrases, to conjoin them, translating to English as «both … and». An example of this from outside the corpus is given in (10).

(10) *Атам да, апам да каршы болду.*
'Both my father and my mother were against it.'

The meanings and distributions of many of these uses are similar. While the first use has a very distinct meaning and distribution, meanings 2 and 3 are very similar, as are 3 and 4; additionally, meaning 5 seems like it could be understood as a repeated use of 4, or possibly 3.

Benli [Benli, 2023] analyses да in all uses as a coordinating conjunction (CCONJ), attached to its head with a mark dependency. According to Universal Dependencies guidelines [Zeman, 2023], however, coordinating conjunctions conjoin two syntactic constituents with no subordination relationship, and mark is the dependency for a word that is used to subordinate one clause to another. Neither of these types of relationships hold in any of these examples.

In Kazakh, the first use of да does not exist, and an additional use to conjoin two parallel constituents is found *(e.g., Астана елімізге қайырлы да құтты қала болды).* As for the remaining uses, Tyers & Washington [Tyers, 2015] and Makazhanov et al. [Makazhanov, 2015] mostly annotate these as ADV, with an advmod dependency on not the preceding element, but the root. For example, in (4), да 'also' would be an advmod dependency on *болду* 'was', as opposed to *атам* 'my father'. Given the analysis as an adverb, this dependency attachment is somewhat sensible, as the general guidelines for advmod state that it indicates modifier to a predicate or modifier word.

However, the subtype advmod:emph appears to be dedicated specifically to indicating an intensifier or emphasising word that can modify various parts of speech, including nouns and prepositional

phrases. This dependency relation is used in Tatar [Taguchi 2022] and Turkish[1] treebanks for uses similar to *да*.

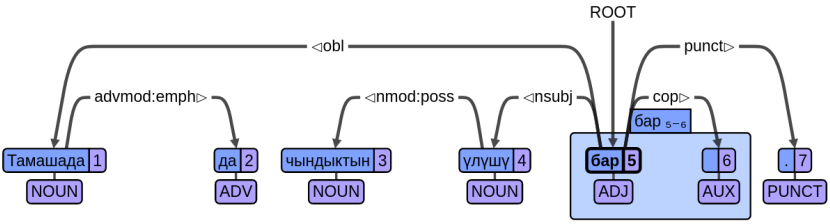
A Kazakh word with a similar distribution to the first use of да (although different meaning), *ғой/қой* (corresponding to Kyrgyz *го*), is annotated by Tyers & Washington [2015] and Makazhanov et al. [2015] as PART, with a discourse relation to the root.

Given all of this, we opt to analyse *да* in the following ways:

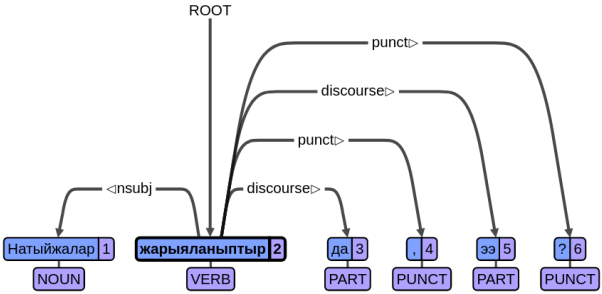
1. Post-predicate modal particle *да* is PART, with a discourse relation to the root.

2-5. Intensifier / emphasis uses of *да* are ADV, with an advmod:emph relation to the word it intensifies / emphasises.

Examples of this from our corpus are provided in Graphs 5 and 6.



Graph 5: Example of annotation of advmod:emph dependent of NOUN, corresponding to sentence (8).



Graph 6: Example of annotation of discourse да, corresponding to sentence (6).

4.2.2. *эле*

The word *эле*, translating as ‘only, just’ (ignoring copula uses per §4.1), may occur after nearly any part of speech or phrase type in Kyrgyz:

[1] Per UD documentation of the existing four Turkish treebanks: https://universaldependencies.org/tr/dep/advmod-emph.html

• after nouns: *бала эле* ‘just a child’
• after adjectives: *кичинекей эле* ‘not that big’
• after numbers: *эки эле* ‘just two’
• after adverbs: *кечээ эле* ‘just yesterday’
• after adverbial clauses: *үч күн өткөндөн кийин эле* ‘only after three days had passed’

Benli [2023] analyses эле as ADV, with an advmod relation (except in cases like чын эле ‘really’ where it is given a fixed or compound relation), and Tyers & Washington [2015] and Makazhanov et al. [2015] do the same with the Kazakh word *ғана/қана*, which has a similar distribution and meaning. However, the distribution of *эле*, including after nouns and numbers, makes it difficult to consider it a true adverb. However, we feel that like *да*, these uses of *эле* fit the intended use of the dependency relation advmod:emph. Hence, we annotate it this way, along with the POS tag ADV.

4.2.3. *бар* and *жок*

The Kyrgyz words *бар* and *жок* are used in constructions that translate into English roughly as ‘there is/are’ and ‘there is not / are not’, respectively. With either possession or locative morphology, they can translate into ‘has/have’ and ‘do(es) not have’ constructions. Despite these verb-based translation, the fact that these words occur in copula constructions (11) and *be* verbs (12) is strong evidence that they are in fact either adjectives or nouns.

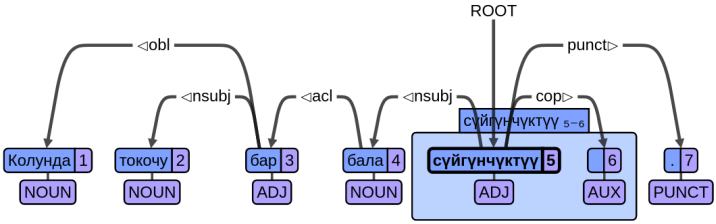
(11) *Сен турганда мен бармын.*
sen tur-GAn-DA men bar-MIn
you stand-VN-LOC I present-COP.NPST.1SG
‘I’m there when you get up.’
(literally: ‘I’m present’)

(12) *Эртең Бишкектин айрым жерлеринде суу жок болот.*
erteŋ Bişkek-NIn ayrım jer-LAr-(s)I(n)-DA suu joq bol-E-t
tomorrow Bishkek-GEN some place-PL-POSS.3-LOC water absent be-NPST-3
‘Tomorrow there will not be water in some places in Bishkek.’
(literally: ‘water will be absent’)

Examples like (8), from the corpus, push us to consider these words adjectives, translating literally as «present» and «absent», respectively.

(13) *Колунда токочу бар бала сүйгүнчүктүү.*
qol-(s)I(n)-DA toqoç-(s)I(n) bar bala süygünçük-LUU.
hand-POSS.3-LOC loaf-POSS.3 present child darling.
'The child with / who has the loaf (of bread) in their hands is darling.'
(literally: '[in their hand their loaf (being) present] child' or 'the child [whose loaf is present in their hand]')

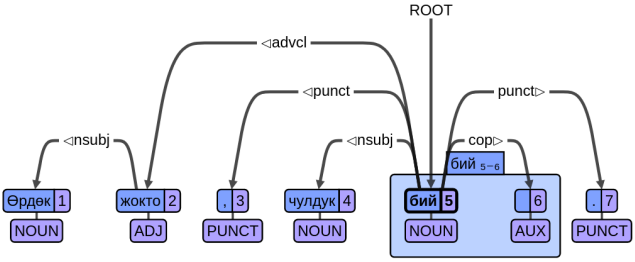
In (13), *бар* is the predicate of nominal subject *токочу* in a sort of copula construction, but the entire phrase is an adjective clause modifying *бала*. This is depicted in Graph 7.



Graph 7: UD dependency graph for sentence (8).

Furthermore, *бар* and *жок* can be used as nouns in Kyrgyz, receiving regular nominal morphology. An example from the corpus is (14), where *жокто* forms the head of an adverbial clause dependent on the main copula construction. A dependency graph for (14) is shown in Graph 8.

(14) *Өрдөк жокто, чулдук бий.*
ördök joq-DA çulduq biy
duck absent-LOC sandpiper bey
'When the duck is absent, the sandpiper is king.'



Graph 8: UD dependency graph for sentence (14).

While *бар* and *жок* are often used in predicates, these previous two examples show their uses in other contexts. We understand these

words to be categorised as adjectives in Kyrgyz no matter what kind of construction they are encountered in.

4.2.4. *керек*

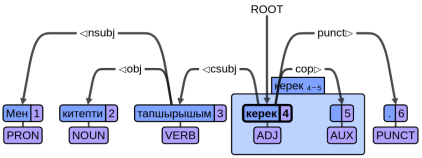
In Kyrgyz the word *керек* is used in ‘need to’ phrases, like that in (15).

(15) *Мен китепти тапшырышым керек.*
men kitep-NIt apşır-(I)ş-(I)m kerek.
I book-ACC turn.in-VN-POSS.1SG needed
‘I need to return the book.’
(literally: ‘me returning the book is needed/necessary.’)

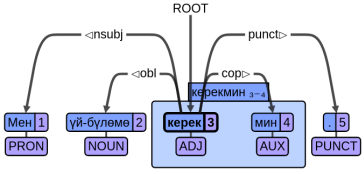
Due to this kind of translation in English as well as the distribution of cognates in some other Turkic languages (e.g., Turkish), it may be tempting to analyse *керек* as a verb, as Benli [2023] mostly does.[1] However, unlike the behaviour of said cognates, *керек* in Kyrgyz does not take any verbal morphology, suggesting that it is not a verb. Instead, it has a morphological and syntactic distribution more like that of a noun or adjective, as in sentences like (16); in this example, it comprises a non-finite predicate, and has a copula morpheme (§4.1) attached to it.

(16) *Мен үй-бүлөмө керекмин.*
men üy-bülö-(I)mA kerek-mIn
I family-POSS.1SG:DAT needed-COP.NPST.1SG
‘My family needs me.’
(literally: ‘I am needed/necessary to my family.’)

As with *бар* and *жок*, we opt to analyse *керек* as an adjective (ADJ) with a literal paraphrase of ‘needed’ or ‘necessary’, although an analysis as a noun (with a reading like ‘a needed/necessary thing’) might also be possible. We annotate these sentences as shown in Graphs 9 and 10, with *керек* as a copular predicate, having a clausal subject (csubj) or nominal subject (nsubj) dependent, respectively.



Graph 9: Dependency graph of sentence (15), showing керек annotated as ADJ with a csubj dependent.



Graph 10: Dependency graph of sentence (16), showing керек annotated as ADJ with an nsubj dependent.

[1] In a few instances Benli (2023) instead analyses *керек* as a NOUN.

4.3. Null-headed clauses

Turkic languages exhibit a number of phenomena where null or empty heads are posited. This term refers to a phrase operating as if a lexical head is present, despite one not being overtly realised. Three different instances of this process are discussed here: substantivsed verbal adjectives (§4.3.1), substantivised relativised locative expressions (§4.3.2), and substantivised genitive expressions (§4.3.3).

4.3.1. Substantivised verbal adjectives

An example of this phenomenon is «substantivised» verbal adjectives [see Washington et al. 2022]. In these constructions, a verbal adjective modifies a noun that is not present, but is understood through the nominal morphology that is in turn attached to the verbal adjective. Verbal adjectives in Turkic are used to form relative clauses, so these may also be considered «headless» relative clauses. These may be read in English as «(the) person/thing/one who/that». Examples of this type of construction are presented in (17) and (18), sentences drawn from the corpus.

(17) *Колуң менен кылганды, мойнуң менен тартасың.*
qol-(I)ŋ menen qıl-GAn-NI moyun-(I)ŋ menen tart-E-sIŋ.
hand-POSS.2SG with make-VADJ-ACC neck-POSS.2SG with pull-NPST-2SG
'You will pull with your neck what you make with your hands.'

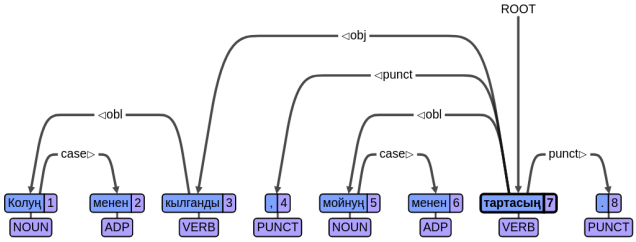
(18) *Балалуу болбогон кубанганды билбеген.*
bala-luu bol-BA-GAn quban-GAn-NI bil-BA-GAn.
child-ORN be-NEG-VADJ be.happy-VN-ACC know-NEG-PST;3
'One who has not had children has not known being happy.'

Both of these sentences could be stated with an additional word added after the verbal adjective suffix and have almost exactly the same meaning, e.g. *кылган нерсени* 'make-VADJ thing-ACC = the thing you make' and *болбогон киши* 'the person who has not had', respectively. Hence, once possibility is to add an additional null node to the UD analysis of these sentences. However, UD standards are strongly against adding null nodes if at all possible.

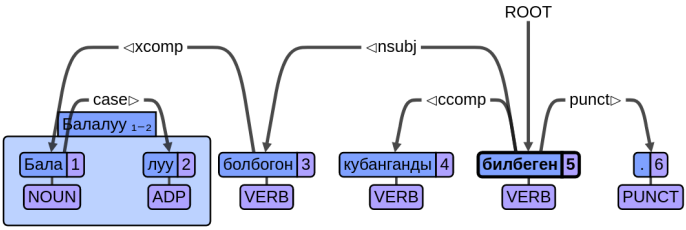
Even without adding null nodes, the most obvious way (to us) of annotating these verbal adjective clauses still shows that are (tacitly) dependent on a nominal head. Specifically, treating them as a nominal object (obj) (17) or nominal subject (nsubj) (18) instead of as a clausal complement (ccomp) (17) or clausal subject (csubj) (18) makes

it clear that the verb is not the head of these phrases. Since they are verbal adjectives (VerbForm=Part), they are not really able to operate as objects or subjects (clausal or nominal) on their own anyway.

This approach is shown in the annotated versions of (17) and (18) in Graphs (11) and (12), respectively.



Graph 11: UD annotated version of sentence (17), without a null head explicitly represented, and showing the empty-headed verbal adjective as an object, as opposed to a clausal complement.



Graph 12: UD annotated version of sentence (18), without a null head explicitly represented, and showing the empty-headed verbal adjective as a nominal subject, as opposed to as a clausal subject.

Benli [2023] inconsistently treats empty-headed verbal adjectives as verbal nouns (VerbForm=Vnoun) and verbal adjectives (VerbForm=Part). In nearly every instance that they were analysed as verbal nouns, they include plural morphology, which is not semantically compatible with verbal nouns in most Turkic languages. Tyers & Washington (2015) treat these forms fairly consistently as gerunds (VerbForm=Ger), equivalent to verbal nouns, and as clausal dependents instead of nominal dependents.

4.3.2. Substantivised relativised locative expressions

The locative case in Kyrgyz is *-DA*, and can only be used adverbially. A derived form of the locative case used attributively is *-DAGI*. An example from corpus is presented in (19).

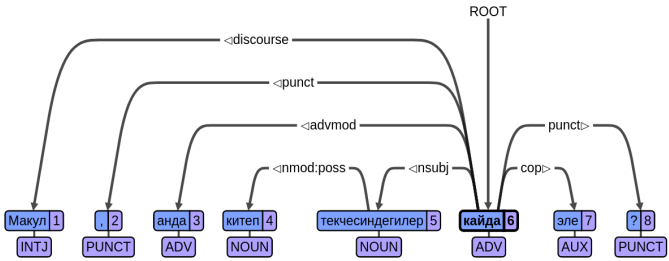
(19) *Алыстагы душмандан аңдып жүргөн дос жаман.*
alıs-DAGI duşman-DAn aŋdı-(I)p jür-GAn dosjaman.
far-LOC;ATTR enemy-ABL spy-INF go.around-VADJ friend bad.
'A friend who spies on you is worse than an enemy who is far away.'

Here *алыстагы* is an nmod:loc dependent on the noun *душман*.

Forms in -DAGI can also have an empty head, and hence can function as nominal heads and take nominal morphology. An example from the corpus is presented in (20).

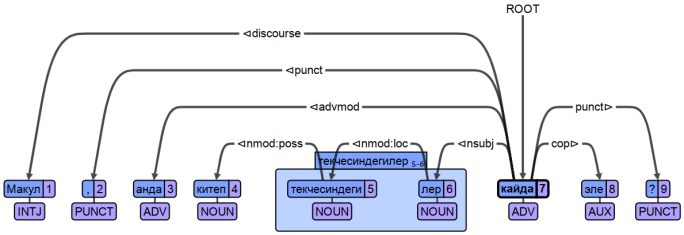
(20) *Макул, анда китеп текчесиндегилер кайда эле?*
maqul anda kitep tekçe-(s)In-DAGI-LAr qayda ele?
okay then book shelf-POSS.3-LOC;ATTR-PL where were?
'Okay, then where were the ones on the bookshelf?'

In one UD annotation of this sentence, presented in Graph 13, the subject misleadingly appears to simply be an inflected form of *китеп текчеси* 'bookshelf', despite the existence of another participant, hidden from the analysis due to it not having a surface realisation. Additionally, in morphological features, there are two distinct items: a singular bookshelf, and a plural set of items on the shelf–which number to annotate this form with is not clear according to UD guildelines.



Graph 13: A UD annotation of sentence (20), without an extra token for the additional «empty» participant.

The only other way to annotate such structures, as we see it, would be to break the problematic form into two subtokens, as in Graph 14. It is not clear to us that this is preferable, but it solves the issue of associating features for multiple participants with one form. This approach also clarifies that there are multiple participants. For now, we have gone with this approach in the corpus.



Graph 14: A UD annotation of sentence (20) with an extra token for the additional «empty» participant.

4.3.3. Substantivised genitive expressions

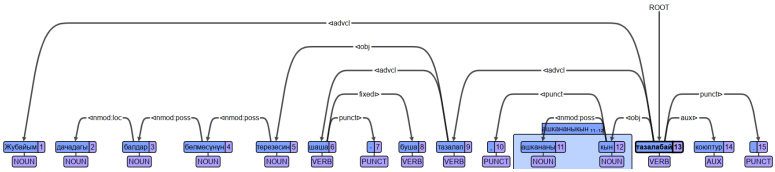
Also wrapped up in discussions of «*ki*» in Turkish are substantivised genitive expressions. In Kyrgyz these are formed with *-NIKI*. Despite the similarity in Turkish and potential etymological unity of the *KI* element, the way *-NIKI* works is conceptually different from *-DAGI*: the former creates a substantivised form of the genitive, whereas the latter creates an attributive form of an adjective, and, on occasion, a substantivised form as discussed in §4.3.2. An example of Kyrgyz *-NIKI* is presented in sentence (21), which is drawn from our corpus.

(21) *Жубайым дачадагы балдар бөлмөсүнүн терезесин*
jubay-(I)m daça-DAGI bala-LAR bölmö-(s)I(n)-nIn tereze-(s)In
spouse-POSS.1SG summer.house-LOC.ATTR child-PL room-POS
S.3-GEN window-POSS.3:ACC

шаша-буша таз лап, ашкананыкын тазалабай коюптур.
şaş-E-buş-E tazala-(I)p, aşqana-NIKIn tazala-BA-E qoy-(I)
ptIr.
rush-VADV-rush-VADV clean-VADV kitchen-GEN.SUBST:ACC
clean-NEG-INF put-PST.IDF;3

'My spouse cleaned the windows of the children's room at the summer house hastily, but didn't clean the kitchen's.'

Our approach to dealing with *-NIKI* is the same as for substantivised forms of *-DAGI*: we explicitly add an empty head to our dependency graph, and give it the features of the head of the phrase, with an nmod:poss relationship to the noun with *-NIKI* morphology. We segment the subtokens in the middle of the morpheme, e.g. ашкананыкын, where we treat the first part as a genitive (the full equivalent of which is ашканандын) and the second part consists of *KI* and any additional morphology–here, accusative case. A dependency graph showing this is provided in Graph 15.



Graph 15: Dependency graph for sentence (21), with the null head a separate token.

Tyers and Washington [Tyers, 2015] parse cognate forms as two separate tokens, but make the second one a case dependent on the first, preventing the analysis of two separate participants in the event. Benli [Benli, 2023] mysteriously treats some *-NIKI* forms as being ADJ and amod dependents; however, one form in the corpus is treated as having two case markings, one being accusative (instead of genitive) and the other being the case of the empty head.

4.4. Inflection versus derivation

A number of suffixes that occur with nouns in Kyrgyz are not commonly analysed as case suffixes among Kyrgyz linguists. For example *-LUU* could be considered an ornative case suffix, but it is not. Similarly, *-sIz* could be considered abessive or privative case, *-DAy* could be considered semblative case, and *-çA* could be considered an adverbial case.

There are at least three different ways to analyse nouns with such morphology:

1. One way is to treat the nouns as deriving adjectives or adverbs; e.g., *балалуу* in (18) could be treated simply as an adjective ‘having child(ren)’ that happens to be derived from the noun *бала* ‘child’. This analysis is unsatisfying because these morphemes act quite productively.

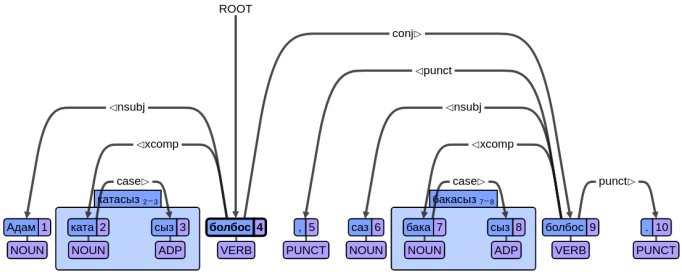
2. Another possible analysis is to analyse these morphemes as case marking, as is done for other cases in Turkic. So, the lemma of *балалуу* would be *бала*, and there would be annotation for ornative case in the grammatical features. One disadvantage of this approach is that unlike other case morphemes, these morphemes both do not enable the noun to act as a core argument of a verb (i.e., one that is involved in case demotion or promotion related to grammatical voice inflection), and can be used to create attributive (adjective-like) forms. Additionally, the names for these cases are not standardised in Kyrgyz grammar, and

the names that seem the most obvious to use are not all standard within Universal Dependencies.

3. These morphemes can be treated as cliticised postpositions. This has the advantage of highlighting their productivity and their difference from case suffixes. The main disadvantage of this approach is that tokenisation then does not line up with spaces. Another disadvantage is that the other case morphemes could also be annotated this way, and not doing so introduces some level of arbitrariness into the corpus.

We decided to go with the third approach for the most part, as shown in (22) and Graph 16, depicting a sentence drawn from our corpus.

(22) *Адам<u>катасыз</u> болбос, саз <u>бакасыз</u> болбос.*
adam <u>qata-sIz</u> bol-BAs, saz <u>baqa-sIz</u> bol-BAs.
person <u>error-ABE</u> be-NEG;FUT.IDF marsh <u>frog-ABE</u> b e - NEG;FUT.IDF
'A person won't be <u>without errors</u>, a marsh won't be <u>without frogs</u>.'



Graph 16: Annotated version of sentence (22), showing case-like morphology treated as postpositions.

Benli [Benli, 2023] uses a mixture of the first two approaches. For example, *-LUU* and *-sIz* forms are sometimes annotated as NOUN, other times as ADJ, and other times as ADV, and often with the lemma being the noun lemma which the morpheme is attached to, regardless of the part of speech annotated. Tyers and Washington [Tyers, 2015] use a mixture of the first and third approaches; for example, баласыз 'without children' is treated as having two subtokens, with сыз an ADP, while сансыз 'without count' is treated simply as an adjective.

5. Conclusion

In this paper we have presented several issues of syntactic annotation relevant to the annotation of a forthcoming Universal Dependencies

annotated corpus of Kyrgyz text. We have compared these issues to the existing UD corpora of Kazakh [Tyers and Washington, 2015; Makazhanov et al., 2015] and Kyrgyz [Benli 2023]. We have weighed the advantages and disadvantages of various approaches, and argue for specific solutions to these issues. Compared to the existing Kyrgyz treebank [Benli 2023], we aim to present a more comprehensive analysis of ongoing annotation issues and to build a treebank of larger size and domain coverage.

The UD Kyrgyz corpus presented here significantly contributes to the syntactic resources available for the Kyrgyz language. This corpus will serve as a valuable resource for studying the syntax and grammatical structure of the Kyrgyz language, as well as for developing language technologies such as dependency parsers and machine translation systems. It provides higher-quality annotated data compared to the previously available UD Kyrgyz treebank [Benli 2023], addressing the need for expanded Kyrgyz resources to support natural language processing applications. The inclusion of a new syntactic corpus in the Universal Dependencies framework for Kyrgyz will not only enhance the quality of linguistic research but also contribute to the broader goal of enhancing the representation of underrepresented languages in language technology. By addressing existing limitations and inaccuracies, this endeavour enables the Kyrgyz language to be better understood, studied, and utilized in various language-related applications along with promoting the development of resources to support Kyrgyz natural language processing.

Acknowledgments

We gratefully acknowledge the foundational work of Prof. Elke Teich and MSc. Jörg Knappen at the Universität des Saarlandes in developing the initial Manas-UdS Kyrgyz corpus that provided much of the textual data for this UD project. Their efforts in compiling a representative sample of texts across various domains helped ensure a strong underlying syntactic corpus. We are also grateful to the participants of the 2023 UD Turkic Workshop, who offered insightful feedback and discussion. We also sincerely thank our talented students Aidai Abitova, Alina Iskenderova, Alina Nijazbekova, Azima Naamatbekova, Bermet Ulukbekova, Cholpon Kultaeva, Kurmanjan Ydyrysova, Meerim Taalaibekova, Suyun Tostonova, and Zuura Mirlanova at Kyrgyz-Turkish Manas University who diligently performed a first pass of syntactic annotation for many of the Kyrgyz sentences according to the Universal Dependencies guidelines. The students' careful application of UD principles during the annotation process, while also thoughtfully handling

ambiguities and inconsistencies, resulted in a high-quality base for the resource under development. Their contributions have made the annotated Kyrgyz UD corpus a valuable asset for future research and tool development for the natural language processing community working with the Kyrgyz language.

REFERENCES

1. Aili M., Mushajiang W., Yibulayin T., Yan Liu K. A. In Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016). Osaka, Japan. 2016. Pp. 44–50.

2. Benli İ. UD_Kyrgyz-KTMU: Universal Dependency treebank for Kyrgyz. 2023 https://github.com/UniversalDependencies/UD_Kyrgyz-KTMU, https://universaldependencies.org/treebanks/ky_ktmu/index.html.

3. Çetinoğlu Ö., Çöltekin Ç. Two languages, one treebank: building a Turkish-German code-switching treebank and its challenges. In: Language Resources and Evaluation. 2023. Vol. 57, pp. 545–579. https://doi.org/10.1007/s10579-021-09573-1.

4. Hakkani-Tur D., Oflazer K., Tur Go. Statistical Morphological Disambiguation for Agglutinative Languages. Computers and the Humanities. 2002. Vol. 36. 381-410. http://doi.org/10.1023/A:1020271707826.

5. Imanalieva Zh. Кыргыз жана орус тилдеринде синтаксистик катыштардын синтаксистик өзгөчөлүктөрү, Бишкек, 2015. С. 197–200. http://www.science-journal.kg/media/Papers/nntiik/2015/11/197-200.pdf (Accessed: 20 October 2023).

6. Kasieva A., Knappen J., Fischer S., and Teich E. A new Kyrgyz corpus: sampling, compilation, annotation. Poster at: 42. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, Hamburg (Germany), March 2020. https://www.zfs.uni-hamburg.de/dgfs2020/programm/abstracts/dgfs2020-clp-kasieva.pdf, https://corpora.clarin-d.uni-saarland.de/cqpweb/kyrgyz_2022_03_08.

7. Makazhanov A., Sultangazina A., Makhambetov O., and Yessenbayev Zh. Syntactic Annotation of Kazakh: Following the Universal Dependencies Guidelines. A report. In: Proceedings of the 3rd International Conference on Computer Processing in Turkic Languages (TurkLang 2015). Kazan, Tatarstan, 2015. Pp. 338–350. http://www.turklang.org/en/turklang-2015-2/.

8. Nivre J., Marneffe M., Ginter F., Goldberg Y., Hajic J., Manning Ch., McDonald R., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., Zeman D. Universal Dependencies v1: A Multilingual Treebank Collection. In Proc. of LREC 2016. Pp. 1659-1666.

9. Oflazer K., Say, B., Zeynep, D., Tur, G. Building a Turkish Treebank. Abeillé, 2003. http://doi.org/10.1007/978-94-010-0201-1_15.

10. Sulubacak U., Gokirmak M., Tyers F., Çöltekin Ç., Nivre J., Eryiğit G. Universal Dependencies for Turkish. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 3444–3454, Osaka, Japan. 2016. https://aclanthology.org/C16-1325.

11. Taguchi Ch. UD Tatar-NMCTT: Universal Dependency corpus for Tatar. 2022. https://github.com/UniversalDependencies/UD_Tatar-NMCTT, https://universaldependencies.org/treebanks/tt_nmctt/index.html.

12. Thompson A. Syntactic Parallelism and Structure in Kyrgyz Proverbs (Bachelors thesis). Bryn Mawr College, Pennsylvania. 2021.

13. Tyers F., Sheyanova M., Washington J. UD Annotatrix: An annotation tool for Universal Dependencies. In: Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT). Praha, Česko, 2017. Pp. 10–17. https://aclanthology.org/W17-7604.

14. Tyers F., Washington J. Towards a free/open-source universal-dependency treebank for Kazakh. In: Proceedings of the 3rd International Conference on Computer Processing in Turkic Languages. TurkLang 2015. Kazan, Tatarstan. 2015 Pp. 276–289. http://www.turklang.org/en/turklang-2015-2/.

15. Tyers F., Washington J., Çöltekin Ç., Makazhanov A. An assessment of Universal Dependency annotation guidelines for Turkic languages". In: Proceedings of the Fifth International Conference on Turkic Language Processing. TurkLang 2017. Vol. 1. Pp. 276–297. http://www.turklang.org/en/turklang-2017-2/.

16. Washington J., Ipasov M., Tyers F. A finite-state morphological transducer for Kyrgyz. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA). Istanbul, Turkey. 2012. Pp. 934–940 https://aclanthology.org/L12-1642/.

17. Washington J., Salimzyanov I., Tyers F. Finite-state morphological transducers for three Kypchak languages. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA). Reykjavik, Iceland. 2014. Pp. 3378–3385, https://aclanthology.org/L14-1143/.

18. Washington J., Tyers F., Salimzianov I. Non-finite verb forms in Turkic exhibit syncretism, not multifunctionality. In: Shagal, Ksenia, Pavel Rudnev, and Anna Volkova (eds.), Folia Linguistica, vol. 56, no. 3, Special Issue: Multifunctionality and syncretism in non-finite forms, 2022. Pp. 693–742. https://doi.org/10.1515/flin-2022-2045.

19. Zeman D. et al. Universal Dependencies 2.12». In: LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. 2023 http://hdl.handle.net/11234/1-5150.

20. Джумалиева Г.К., Касиева А.А., Мусажанова С.Ж. [Dzhumalieva G.K., Kasieva A.A., Musazhanova S.J.]. Адаптация терминов веб-проекта универсальные зависимости на кыргызский язык [Adaptation of Web Project Terms for Universal Dependencies in the Kyrgyz Language]. In: Вестник КРСУ [Bulletin of KRSU]. Bishkek, 2023. Vol. 23, № 6, pp. 71–75. http://doi.org/10.36979/1694-500X-2023-23-6-71-75.

21. Мусажанова С. Ж., Касиева А. А., Джумалиева Г. К. [Musazhanova S. J., Kasieva A. A., Dzhumalieva G. K.]. Синтаксическая аннотация кыргызского языка на основе новосозданного корпуса [Syntactic Annotation of the Newly-Created Kyrgyz Corpus]. Вестник Иссык-Кульского университета [Bulletin of the Issyk-Kul University], Karakol, 2023. №54. Pp. 140–148.