Article/Мақала/Статья
IRSTI 16.31.21 https://doi.org/10.55491/2411-6076-2025-2-153-167

Meerim Ryspakova[1*], Aigul Tursunova[2]
[1*]Corresponding author, Doctoral student, I. Arabaev Kyrgyz State University, Kyrgyzstan, Bishkek,
ORCID: 0000-0001-5689-1365 E-mail: meerim.ryspakova@gmail.com
[2]Doctoral student, I. Arabaev Kyrgyz State University, Kyrgyzstan, Bishkek,
ORCID: 0000-0003-4483-2581 E-mail: aygulya2380@gmail.com

"UNIVERSAL DEPENDENCIES" FOR SYNTACTIC ANALYSIS OF THE KYRGYZ LANGUAGE: CURRENT STATE AND PROSPECTS

Abstract. Kyrgyz, a Turkic language with over 4.4 million speakers concentrated primarily in Kyrgyzstan and adjacent regions of Central Asia, faces a significant disparity in computational linguistic resources compared to languages with similar or even smaller speaker populations. Despite its status as a government language and cultural cornerstone, Kyrgyz remains underrepresented in the digital linguistic landscape. This investigation examines the application of the Universal Dependencies (UD) framework – an annotation system engineered to facilitate cross-linguistic syntactic comparability – to the structural complexities of Kyrgyz. We endeavor to identify optimal annotation strategies that faithfully represent Kyrgyz-specific syntactic phenomena while adhering to the principled constraints of the UD paradigm. The establishment of standardized syntactic resources for Kyrgyz carries dual significance: it advances linguistic typology by incorporating data from an underrepresented language family, while simultaneously laying groundwork for practical natural language processing applications crucial for Kyrgyz speakers' participation in the digital sphere. Our methodological approach encompasses rigorous analysis of nascent Kyrgyz treebanks, comparative evaluation of annotation strategies employed for genetically related Turkic languages, and systematic examination of four fundamental annotation challenges: the representation of Kyrgyz's defective copula system, the classification of multifunctional grammatical particles, the annotation of constructions with implicit heads, and the demarcation between inflectional and derivational morphology in this highly agglutinative language. Our analysis reveals that achieving the dual objectives of linguistic fidelity and cross-linguistic consistency necessitates judicious adaptation of UD guidelines to accommodate Kyrgyz-specific structures. We advance unified annotation solutions that preserve the integrity of Kyrgyz linguistic patterns while facilitating meaningful cross-linguistic comparison. This research not only contributes substantively to computational resources for Kyrgyz but also establishes annotation principles with broader applicability to typologically similar agglutinative languages. The practical implications extend to enhanced guidelines for Kyrgyz treebank development, which will consequently improve parser accuracy and catalyze the development of essential language technology tools for Kyrgyz speakers.

Keywords: Kyrgyz language; Universal Dependencies; syntactic annotation; treebanks; Turkic languages; computational linguistics

For citation: Ryspakova, M., Tursunova, A. "Universal Dependencies" for Syntactic Analysis of the Kyrgyz Language: Current State and Prospects. *Tiltanym*, 2025. No. 2 (98). P. 153-167.

DOI: https://doi.org/10.55491/2411-6076-2025-2-153-167

Мээрім Рыспақова[1*], Айгүл Тұрсұнова[2]
[1*]автор-корреспондент, докторант, И. Арабаев атындағы Қырғыз мемлекеттік университеті,
Қырғызстан, Бішкек қ., ORCID: 0000-0001-5689-1365 E-mail: meerim.ryspakova@gmail.com
[2]докторант, И. Арабаев атындағы Қырғыз мемлекеттік университеті,
Қырғызстан, Бішкек қ., ORCID: 0000-0003-4483-2581 E-mail: aygulya2380@gmail.com

ҚЫРҒЫЗ ТІЛІН СИНТАКСИСТІК ТАЛДАУ ҮШІН «ӘМБЕБАП ТӘУЕЛДІЛІКТЕР»: ҚАЗІРГІ ЖАҒДАЙЫ ЖӘНЕ БОЛАШАҒЫ

Аңдатпа. Қырғыз тілі – түркі тілдер тобына жататын, негізінен Қырғызстан мен Орталық Азияның іргелес аймақтарында шоғырланған 4,4 миллионнан астам сөйлеушісі бар қырғыз тілі тіл иелерінің саны ұқсас немесе одан да аз тілдермен салыстырғанда есептеуіш лингвистикалық ресурстарда айтарлықтай теңсіздікке тап болып отыр. Мемлекеттік тіл және мәдени тірек ретіндегі мәртебесіне қарамастан, қырғыз тілі цифрлық лингвистикалық ландшафтыда әлі де жеткіліксіз қамтылған. Бұл зерттеу Әмбебап тәуелділіктер шеңберін (Universal Dependencies, UD) – тіларалық синтаксистік салыстыруды қамтамасыз ету үшін әзірленген аннотация жүйесін қырғыз тілінің құрылымдық ерекшеліктеріне қолдануды қарастырады. Біз UD парадигмасының негізгі шектеулерін сақтай отырып, қырғыз тіліне тән синтаксистік құбылыстарды шынайы көрсететін оңтайлы аннотация стратегияларын анықтауға тырысамыз. Қырғыз тілі үшін стандартталған синтаксистік ресурстарды құру екі жақты маңызға ие: ол жеткіліксіз ұсынылған тіл тобынан деректерді қосу арқылы лингвистикалық типологияны алға жылжытады және сонымен бірге қырғыз тілінде сөйлейтіндердің цифрлық салаға қатысуы үшін маңызды табиғи тілді өңдеуші қолданбалар үшін негіз

қалайды. Біздің әдіснамалық тәсіліміз жаңадан пайда болған қырғыз тілінің синтаксистік корпустарын мұқият талдауды, генетикалық жақын түркі тілдеріне қолданылатын аннотация стратегияларын салыстырмалы бағалауды және төрт негізгі аннотация мәселелерін жүйелі зерттеуді қамтиды: қырғыз тілінің ақаулы көмекші етістік жүйесін көрсету, көп функциялы грамматикалық бөлшектерді жіктеу, жасырын негізгі элементтері бар конструкцияларды аннотациялау және осы жоғары агглютинативті тілдегі сөз түрлендіруші және сөзжасамдық морфологияның аражігін ажырату. Біздің талдауымыз лингвистикалық дәлдік пен тілдер арасындағы сәйкестіктің қос мақсаттарына жету үшін қырғыз тіліне тән құрылымдарға UD нұсқауларын шебер бейімдеу қажет екенін көрсетеді. Біз қырғыз лингвистикалық үлгілерінің тұтастығын сақтай отырып, тілдер арасында мағыналы салыстыруға мүмкіндік беретін бірыңғай аннотация шешімдерін ұсынамыз. Бұл зерттеу қырғыз тілі үшін есептеу ресурстарына елеулі үлес қосып қана қоймай, типологиялық ұқсас агглютинативті тілдерге кеңінен қолданылатын аннотация принциптерін де белгілейді. Практикалық салдарлар қырғыз синтаксистік корпустарды дамыту үшін жетілдірілген нұсқауларды қамтиды, бұл өз кезегінде парсер дәлдігін жақсартып, қырғыз тілінде сөйлейтіндер үшін маңызды тілдік технологиялық құралдарды әзірлеуді жеделдетеді.

Тірек сөздер: қырғыз тілі; Әмбебап тәуелділіктер; синтаксистік аннотация; синтаксистік корпустар; түркі тілдері; компьютерлік лингвистика

Сілтеме жасау үшін: Рыспақова М., Тұрсұнова А. Қырғыз тілін синтаксистік талдау үшін «Әмбебап тәуелділіктер»: қазіргі жағдайы және болашағы. *Tiltanym,* 2025. №2 (98). 153-167-бб. (ағыл. тілінде)

DOI: https://doi.org/10.55491/2411-6076-2025-2-153-167

Мээрим Рыспакова[1*], Айгүл Турсунова[2]

[1*]автор-корреспондент, докторант, Кыргызский государственный университет им. И. Арабаева, Кыргызстан, г. Бишкек, ORCID: 0000-0001-5689-1365 E-mail: meerim.ryspakova@gmail.com

[2]докторант, Кыргызский государственный университет им. И. Арабаева, Кыргызстан, г. Бишкек, ORCID: 0000-0003-4483-2581 E-mail: aygulya2380@gmail.com

«УНИВЕРСАЛЬНЫЕ ЗАВИСИМОСТИ» ДЛЯ СИНТАКСИЧЕСКОГО АНАЛИЗА КЫРГЫЗСКОГО ЯЗЫКА: ТЕКУЩЕЕ СОСТОЯНИЕ И ПЕРСПЕКТИВЫ

Аннотация. Кыргызский язык, принадлежащий к тюркской языковой семье и насчитывающий более 4,4 миллиона носителей, сосредоточенных преимущественно в Кыргызстане и прилегающих регионах Центральной Азии, сталкивается со значительным неравенством в вычислительных лингвистических ресурсах по сравнению с языками, имеющими сходную или даже меньшую численность носителей. Несмотря на статус государственного языка и культурной основы, кыргызский язык остается недостаточно представленным в цифровом лингвистическом ландшафте. Данное исследование рассматривает применение фреймворка Универсальных Зависимостей (Universal Dependencies, UD) – системы аннотирования, разработанной для обеспечения межъязыковой синтаксической сопоставимости – к структурным особенностям кыргызского языка. Мы стремимся определить оптимальные стратегии аннотирования, которые достоверно отражают специфические для кыргызского языка синтаксические явления, соблюдая при этом принципиальные ограничения парадигмы UD. Создание стандартизированных синтаксических ресурсов для кыргызского языка имеет двойное значение: оно продвигает лингвистическую типологию, включая данные из недостаточно представленной языковой семьи, и одновременно закладывает основу для практических приложений обработки естественного языка, критически важных для участия носителей кыргызского языка в цифровой сфере. Наш методологический подход включает тщательный анализ новых кыргызских синтаксических корпусов, сравнительную оценку стратегий аннотирования, применяемых для генетически родственных тюркских языков, и систематическое исследование четырех фундаментальных проблем аннотирования: представление дефективной системы связок кыргызского языка, классификацию многофункциональных грамматических частиц, аннотирование конструкций с имплицитными главными элементами и разграничение между словоизменительной и словообразовательной морфологией в этом высоко агглютинативном языке. Наш анализ показывает, что достижение двойных целей лингвистической точности и межъязыковой согласованности требует разумной адаптации руководящих принципов UD для размещения специфических для кыргызского языка структур. Мы предлагаем унифицированные решения по аннотированию, которые сохраняют целостность кыргызских лингвистических моделей, одновременно способствуя значимому межъязыковому сравнению. Это исследование не только вносит существенный вклад в вычислительные ресурсы для кыргызского языка, но и устанавливает принципы аннотирования с более широким применением к типологически схожим агглютинативным языкам. Практические последствия включают в себя улучшенные рекомендации для разработки кыргызских синтаксических корпусов, что, в свою очередь, повысит точность парсера и ускорит разработку важных инструментов языковых технологий для носителей кыргызского языка.

Ключевые слова: кыргызский язык; Универсальные зависимости; синтаксическое аннотирование; синтаксические корпуса; тюркские языки; компьютерная лингвистика

Для цитирования: Рыспакова М., Турсунова А. «Универсальные зависимости» для синтаксического анализа кыргызского языка: текущее состояние и перспективы. *Tiltanym,* 2025. №2 (98). С. 153-167. (на англ. яз.)

DOI: https://doi.org/10.55491/2411-6076-2025-2-153-167

Introduction

Language technology has transformed how we interact with information, yet this transformation has been uneven across the world's languages. While speakers of major languages like English, Spanish, or Chinese benefit from robust computational tools enabling everything from machine translation to voice assistants, millions of speakers of languages like Kyrgyz find themselves on the wrong side of the "digital language divide" (Kornai, 2013). This gap is not merely technical but has profound implications for cultural preservation, educational access, and economic opportunity in an increasingly digital world.

The Universal Dependencies (UD) project represents a significant initiative to address this imbalance by creating a standardized framework for syntactic annotation across diverse languages. By establishing consistent guidelines, UD enables both linguistic research and the development of natural language processing applications for previously under-resourced languages. For Kyrgyz and other Turkic languages with rich agglutinative morphology and distinctive syntactic patterns, adapting these guidelines presents unique challenges that require careful consideration of linguistic features not typically encountered in Indo-European languages.

Kyrgyz, an official language of Kyrgyzstan with approximately 4.4 million speakers across Central Asia, currently has limited computational resources compared to languages with similar or even smaller speaker populations. The language currently has only one small UD treebank containing 781 sentences (Benli, 2023), with a second, more comprehensive treebank under development (Kasieva et al., 2023). These initial efforts have revealed several areas where standard UD guidelines require thoughtful adaptation to accommodate Kyrgyz linguistic structures.

The agglutinative nature of Kyrgyz, where complex words are formed through the sequential addition of morphemes to stems, creates particular challenges for tokenization and dependency representation within the UD framework. Features such as null-headed clauses, case-like derivational suffixes, and multifunctional particles don't fit neatly into UD's categories, which were initially developed with Indo-European languages in mind. Similarly, the defective copula system in Kyrgyz presents annotation challenges not encountered in languages with more typical copular verbs.

Materials and methods

In this paper, we examine these challenges through the lens of both linguistic theory and practical implementation, proposing solutions that balance descriptive accuracy with cross-linguistic consistency. Our goal is to contribute to the advancement of Kyrgyz language resources and support the development of natural language processing applications for this important but underresourced language, while also providing insights that may be valuable for the annotation of other Turkic and typologically similar languages.

The development of robust UD resources for Kyrgyz is not merely an academic exercise. It has potential applications in machine translation, information retrieval, educational technology, and digital preservation of cultural heritage. By addressing the specific challenges encountered in Kyrgyz annotation, we contribute to the broader goal of making language technology more inclusive and representative of linguistic diversity.

Literature review

The application of Universal Dependencies to Turkic languages has progressed unevenly over the past decade, with significant variation in both the quantity and maturity of available resources. This uneven development reflects broader patterns in the allocation of research attention and funding across languages, but it also provides a valuable comparative perspective for work on Kyrgyz.

Turkish, with its relatively large speaker population (approximately 88 million worldwide) and economic importance, has received the most attention, with multiple treebanks now available (Sulubacak et al., 2016; Çöltekin et al., 2022). The IMST treebank, converted from an earlier dependency formalism, contains over 5,600 sentences, while the BOUN treebank offers nearly 9,800 sentences with broader domain coverage. These resources have been instrumental in developing parsing tools and other applications, but their approaches to specific syntactic structures don't always transfer well to other Turkic languages due to structural differences and divergent annotation choices.

Kazakh, which is closely related to Kyrgyz and shares many structural features, has an established treebank (Tyers et al., 2015; Washington et al., 2015; Makazhanov et al., 2015) containing approximately 1,078 sentences. This resource has provided valuable insights for Kyrgyz annotation, particularly in

handling constructions common to both languages. Additional Turkic languages with UD resources include Tatar (Taguchi, 2022) with 148 sentences, Uyghur (Aili et al., 2018) with 3,456 sentences, and Yakut (Merzhevich et al., Gerardi et al., 2022) with 299 sentences, each contributing to our understanding of how UD can be applied across this language family.

The diversity of approaches within these treebanks poses both opportunities and challenges. As Tyers et al. (2017) noted in their assessment of UD guidelines for Turkic languages, while there are areas of cross-linguistic consistency, significant divergences exist between treebanks even for closely related languages. They highlighted several unresolved issues including tokenization practices, methods for distinguishing core arguments, approaches to complex predicates, and treatment of copulas – all of which remain relevant in current work on Kyrgyz.

Recent theoretical work has also contributed important insights for Turkic UD annotation. A study by Washington et al. (2022) examined non-finite verb forms in Turkic languages, demonstrating that these forms exhibit syncretism rather than multifunctionality – a finding with important implications for syntactic annotation. By showing that forms previously analyzed as having multiple grammatical functions actually represent distinct homophonous morphemes with specific functions, this work helps clarify the appropriate treatment of verbal forms that are central to Turkic syntax.

The most recent contribution to this field is an examination of pronominalised locative expressions across Turkic languages by Washington et al. (2023), which directly addresses one of the most challenging construction types for UD annotation. Their comparison of approaches across treebanks highlights the need for greater consistency while respecting language-specific structures.

Kyrgyz Language Resources

Computational resources for Kyrgyz have developed gradually but remain limited compared to many languages with similar numbers of speakers. This situation reflects broader patterns in language technology development, where factors beyond speaker population - including economic resources, research infrastructure, and digital literacy - strongly influence resource availability.

A significant milestone in Kyrgyz computational linguistics was the creation of a finite-state morphological transducer by Washington et al. (2012), which now covers over 15,000 stems and provides the foundation for morphological analysis in syntactic parsing. Originally developed as part of the Apertium machine translation project, this open-source resource has become an essential component of Kyrgyz NLP tools. However, as the authors themselves noted, the transducer requires further extension to fully handle complex verbal constructions, derivational morphology, and other phenomena relevant to syntactic analysis.

The first dependency-annotated corpus of Kyrgyz emerged from Thompson's (2021) thesis analyzing syntactic structure and parallelism in Kyrgyz proverbs. Though small in scale, focusing on 85 proverbs, this dataset provided valuable insights into the application of dependency grammar to Kyrgyz and contributed to initial Kyrgyz UD resources. The analysis highlighted the relationship between syntactic parallelism and parataxis, a common feature in proverb structure, and demonstrated the feasibility of applying UD annotation to Kyrgyz.

Building on this foundation, Dzhumalieva et al. (2023) explored challenges in syntactic annotation for Kyrgyz within the UD framework, proposing adapted terminology and outlining manual annotation procedures. Their work emphasized the need for Kyrgyz-specific annotation guidelines to address phenomena not fully covered by general UD documentation, while maintaining cross-linguistic compatibility.

The most substantial text resource for Kyrgyz is the Manas-UdS corpus, created through collaboration between Kyrgyz-Turkish Manas University and Saarland University (Kasieva et al., 2020) with approximately 2 million words drawn from literary works and news sources, this corpus provides the raw textual data from which sentences for syntactic annotation are sampled. While relatively small compared to corpora for major languages, it represents a significant achievement given the limited digitized resources previously available for Kyrgyz.

Musazhanova et al. (2023) documented early efforts in syntactic annotation using UD for Kyrgyz, demonstrating that many grammatical categories of Kyrgyz have not yet been fully explored within the UD framework. Their work highlighted significant gaps in computational linguistics for Kyrgyz and established groundwork for future research on annotated corpus development. The paper emphasized that

syntactic annotation requires not only technical knowledge but also deep understanding of Kyrgyz grammatical structures, some of which don't have clear counterparts in the languages for which UD was initially developed.

These resources collectively provide a foundation for Kyrgyz computational linguistics, but significant work remains to develop robust tools comparable to those available for more resourced languages. The development of comprehensive UD treebanks represents an important step in this direction, potentially enabling the creation of parsers, machine translation systems, and other applications that can benefit Kyrgyz speakers.

Results and discussions

Current State of Kyrgyz UD Resources

The landscape of Kyrgyz UD resources is still in its early stages of development, lagging behind many languages with comparable or even smaller speaker populations. This reflects a broader pattern in computational linguistics, where resource development often correlates with factors like economic development, research infrastructure, and digital presence, rather than simply with speaker population.

The first publicly available Kyrgyz UD treebank, UD_Kyrgyz-KTMU (Benli, 2023), contains 781 sentences with 7,451 tokens. The corpus primarily draws on news headlines and selected excerpts from novels and news websites, providing limited genre diversity. While this represents an important first step in Kyrgyz UD development, its modest size and narrow domain coverage restrict its usefulness for comprehensive linguistic analysis or training robust parsers.

To put this in perspective, the smallest Turkish treebank contains over 16,000 tokens, while the largest exceeds 175,000 tokens. Even the Tatar treebank, representing a language with fewer speakers than Kyrgyz, contains 2,280 tokens across 148 sentences. This disparity highlights the significant room for growth in Kyrgyz computational resources.

A detailed examination of the UD_Kyrgyz-KTMU treebank reveals inconsistencies in annotation that reflect the challenges of applying UD guidelines to Kyrgyz. For example, the treatment of copula constructions varies across sentences, with some instances analyzing subject agreement morphemes as features of the subject, while others treat them as separate syntactic elements. Similarly, particles and other "small words" receive inconsistent analyses, sometimes as coordinating conjunctions and sometimes as adverbs, without clear linguistic motivation for the distinction.

These inconsistencies, while understandable in a pioneering resource, create challenges for users of the treebank and highlight the need for more systematic annotation guidelines specific to Kyrgyz. They also underscore the importance of cross-linguistic consistency in annotation practices, particularly for closely related languages like Kyrgyz and Kazakh, where researchers might reasonably expect similar constructions to receive similar analyses.

Developing Resources

A more comprehensive Kyrgyz UD treebank is currently under development (Kasieva et al., 2023), building upon earlier work by Thompson (2021) and incorporating sentences from the Manas-UdS corpus. This new resource currently contains approximately 2,456 tokens across 332 sentences, with samples chosen to represent diverse syntactic constructions rather than being limited to short, simple sentences.

The annotation process for this treebank employs the UD Annotatrix tool (Tyers et al., 2018), which provides validation feedback and supports customization of guidelines for language-specific features. This web-based interface facilitates collaborative annotation and real-time validation, helping ensure compliance with UD structural guidelines while allowing for Kyrgyz-specific adaptations where necessary.

A team of researchers and students at Kyrgyz-Turkish Manas University have performed the annotations, with careful attention to inter-annotator agreement – which exceeded 90% by the completion of the process, suggesting good reliability. This collaborative approach helps ensure both linguistic accuracy and consistency across annotations, although it also introduces the challenge of maintaining consistent practices across a team with varying levels of expertise.

Unlike the existing treebank, this developing resource pays particular attention to challenging syntactic structures like null-headed clauses and copula constructions, with explicit guidelines for their treatment. The annotation team has worked to reconcile linguistic accuracy with UD constraints,

sometimes developing creative solutions for structures that don't fit neatly into the UD framework.

This developing treebank aims to provide more comprehensive coverage of Kyrgyz syntactic phenomena than existing resources. Future iterations plan to expand coverage to include scientific articles, spoken dialogues, and social media text, which will capture a broader range of linguistic features and registers. This domain diversity is crucial for developing NLP applications that can handle real-world language use rather than being limited to formal written text.

The development of this new treebank represents an important step forward for Kyrgyz computational linguistics, but challenges remain. The resource is still small by international standards, and expanding it to a size sufficient for robust parser training will require significant additional effort. Moreover, some fundamental questions about the appropriate treatment of Kyrgyz-specific constructions within the UD framework remain open, as we discuss in the following section.

Key Annotation Challenges

The application of Universal Dependencies to Kyrgyz has revealed several challenging areas that require careful consideration. In this section, we examine four key issues that have emerged during annotation efforts, exploring both their linguistic dimensions and the practical challenges they pose for UD annotation.

Copula Tokenization

Copular constructions, which express equational or attributive relationships without using a full lexical verb, present interesting challenges for syntactic annotation across languages. In Kyrgyz, as in many Turkic languages, these constructions employ strategies that differ significantly from those of Indo-European languages, making their representation within the UD framework particularly challenging.

In non-past tense constructions, Kyrgyz typically expresses the copula through subject agreement morphemes attached directly to the predicate:

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

For past tense constructions, Kyrgyz employs forms of a defective verb э- that appears as a separate orthographic word:

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

The challenge here lies in representing these structurally similar constructions consistently within the UD framework, despite their different surface realizations. The non-past construction involves what appear to be person/number inflections on a non-verbal predicate, while the past tense uses what appears to be a separate auxiliary verb.

Kasieva et al. (2023) propose an elegant solution to this dilemma by treating non-past copula subject agreement morphemes as cliticized forms of the defective copula verb, assigning them the lemma э and POS tag AUX. Under this analysis, both constructions involve the same copula verb, but in non-past forms, it appears as a clitic rather than a separate word. This approach offers several advantages: it creates consistency between the analysis of non-past and past forms, prevents having to assign multiple person/number markers to a single noun, and enables these morphemes to be appropriately labeled as copulas.

From a linguistic perspective, this approach aligns with the understanding that these constructions are functionally equivalent despite their different surface forms. It also reflects the historical development of these morphemes, which likely originated as cliticized forms of the copula verb, even if contemporary speakers may not consciously perceive them as such.

This approach differs significantly from that used by Benli (2023) in the existing Kyrgyz UD treebank, where subject-agreement morphemes on non-verbal predicates are analyzed simply as person features of the subject, sometimes misclassifying them as verbs. Benli also treats forms of *эле* as compound:svc dependents of non-verbal predicates, while analyzing complements of *бол-* forms (which can function similarly to copulas) as amod dependents. This inconsistency creates challenges for users of the treebank and complicates the development of computational tools.

While the solution proposed by Kasieva et al. (2023) introduces some complexity in tokenization, requiring the segmentation of what appears in writing as a single word, it provides greater consistency and more accurately represents the linguistic structure of these constructions. This trade-off between

tokenization complexity and linguistic accuracy is characteristic of many annotation challenges in morphologically rich languages like Kyrgyz.

"Small Words" Annotation

The annotation of grammatically versatile "small words" presents another significant challenge in Kyrgyz UD development. These words often serve multiple functions depending on context and don't fit neatly into the standard part-of-speech categories outlined in UD guidelines. Their appropriate analysis requires careful consideration of both their function in specific contexts and the overall consistency of the annotation scheme.

The word *да* in Kyrgyz serves as a particularly illustrative example of this challenge, with at least five distinct functions identified in corpus analysis:

1. Post-predicate "modal particle" indicating a statement whose truth is considered evident to the interlocutor but is being asserted to explain something else
2. Conditional intensifier, appearing after conditional clauses and roughly equivalent to English "even if"
3. General contrastive intensifier, modifying various phrase types and similar to English "even"
4. General conjoining adverb, comparable to English "also" or "too"
5. Correlative conjunction, used in pairs to express "both ... and"

These functions, while semantically related, involve different syntactic relationships. The first function operates at the discourse level, while the others modify or connect specific elements within the sentence. This functional diversity creates challenges for UD annotation, which requires assigning a single part of speech and dependency relation to each token.

Benli (2023) adopts a one-size-fits-all approach, annotating *да* consistently as a coordinating conjunction (CCONJ) with a mark dependency relation. However, as Kasieva et al. (2023) convincingly argue, this analysis contradicts UD guidelines in two ways: coordinating conjunctions should join constituents without subordination, while the mark relation is reserved for subordinating clauses. Moreover, this uniform treatment fails to capture the functional distinctions between different uses of *да*.

After examining similar particles in other Turkic languages and considering UD practices for comparable elements in other languages, Kasieva et al. (2023) suggest a more nuanced approach: treating the first use (modal particle) as PART with a discourse relation to the root, while classifying the intensifier/emphasis uses (functions 2-5) as ADV with an advmod:emph relation to the modified word. This solution better captures the syntactic behavior of these forms and aligns with UD practices for similar constructions in other languages.

This approach exemplifies a key principle in UD annotation: prioritizing syntactic function over superficial form. By analyzing *да* differently depending on its role in the sentence, rather than assigning it a uniform analysis based on its orthographic form, we create a more linguistically accurate representation of its behavior.

Similar challenges arise with other frequently used words that don't fit neatly into traditional part-of-speech categories. For example, *эле* ("only, just") appears in various contexts and modifies different parts of speech, from nouns to verbs to entire clauses. Kasieva et al. (2023) propose treating it as ADV with an advmod:emph dependency, similar to the intensifier uses of *да*. This analysis captures its function as a modifier while acknowledging its special status as an emphasizer.

The words *бар* and *жок* present a different issue. In most contexts, they translate into English using verbal constructions ("there is/are" and "there isn't/aren't"), which might suggest analyzing them as verbs. However, their behavior in copular constructions and ability to be modified by typical adjective modifiers suggests they function as adjectives with meanings closer to "present" and "absent". By analyzing them as adjectives rather than verbs, we more accurately represent their syntactic behavior in Kyrgyz, even if this creates some divergence from how their translation equivalents might be analyzed in other languages.

The word *керек*, which forms "need to" expressions, is often misanalyzed as a verb due to its English translation. However, in Kyrgyz, it doesn't accept verbal morphology and distributes more like an adjective. Kasieva et al. (2023) analyze it as ADJ with a literal meaning of "needed" or "necessary", taking either clausal subjects (csubj) or nominal subjects (nsubj). This approach prioritizes the word's distribution in Kyrgyz over cross-linguistic translation equivalence, a principle central to sound linguistic

analysis.

These analyses highlight the importance of basing annotations on the syntactic behavior of words in Kyrgyz rather than translation equivalents in other languages. While this may sometimes create apparent divergences in how similar meanings are represented across languages, it results in a more accurate representation of Kyrgyz syntax and aligns with the UD principle of prioritizing syntactic function over semantics in dependency assignment.

Null-headed Clauses

Turkic languages, including Kyrgyz, feature several constructions involving null or empty heads – phrases that function as if they had a lexical head, even though none is overtly present. These constructions pose particular challenges for UD annotation, which generally assumes that each syntactic relationship involves overt lexical items. Null-headed constructions in Kyrgyz require creative adaptation of UD principles to accurately represent their structure while remaining within the framework's constraints.

Substantivized verbal adjectives

In Kyrgyz, verbal adjectives can be "substantivized" – that is, used nominally – when they modify a noun that isn't overtly expressed but is understood through nominal morphology attached to the verbal adjective. These constructions form headless relative clauses, often translated into English with expressions like "the one(s) who..." or "the thing(s) that...":

(3) *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."

In this example, *кылганды* (qıl-GAn-NI, "make-VADJ-ACC") functions as a headless relative clause, referring to something that is made without explicitly naming it. The verbal adjective *кылган* (qıl-GAn, "make-VADJ") would normally modify a noun, but here it stands alone with accusative case marking, indicating that it serves as the direct object of the main verb.

This construction poses a dilemma for UD annotation: should it be treated as a clausal complement (ccomp) because of its verbal origin, or as a nominal object (obj) because of its function in the sentence? The standard UD analysis of headless relative clauses would suggest treating the verbal adjective as the head of a nominal phrase, but this doesn't fully capture the understood relationship between the verbal adjective and an implicit nominal head.

Although UD guidelines discourage adding null nodes to represent missing elements, Kasieva et al. (2023) propose a compromise approach: treating these constructions as nominal objects or subjects rather than clausal complements. This approach indicates that the verb is not the head of these phrases but rather modifies an understood nominal head, better reflecting their syntactic behavior. The verbal adjective's POS is maintained as VERB with the feature VerbForm=Part, but its dependency relation (obj, nsubj, etc.) reflects its nominal function in the sentence.

This solution represents a pragmatic compromise between linguistic accuracy and UD constraints. It captures the nominal function of these constructions while maintaining information about their verbal origin, allowing for both appropriate syntactic analysis and potential recovery of the full structure with an understood head.

Substantivized relativized locative expressions

Kyrgyz uses the locative case suffix *-DA* adverbially, while a derived form *-DAGI* functions attributively. This attributive form can also occur with an empty head, functioning as a nominal:

(4) *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?"

Here, *текчесиндегилер* (tekçe-(s)In-DAGI-LAr, "shelf-POSS.3-LOC.ATTR-PL") refers to items located on the bookshelf without explicitly naming them. The attributive locative suffix *-DAGI* would normally connect a locative phrase to a head noun, but here it appears with plural marking, indicating that it functions as a substantivized expression referring to multiple items.

This construction presents an even more complex challenge for UD annotation than substantivized verbal adjectives. The form contains information about both the location (the bookshelf) and the located items (the understood head), with morphological features (plurality, case) applying to the understood head

rather than the overt noun.

To handle these forms, Kasieva et al. (2023) propose a more radical solution: splitting them into two subtokens, explicitly adding an empty head to the dependency graph and assigning it the features of the understood head of the phrase. The relation between these elements is annotated as nmod:poss, with the noun carrying *-DAGI* morphology as the dependent.

This approach does require some deviation from standard UD tokenization practices, which generally avoid splitting orthographic words except at clitic boundaries. However, it provides the most complete representation of the linguistic structure, allowing both elements – the locative phrase and the understood head – to participate appropriately in syntactic relationships.

Substantivized genitive expressions

Kyrgyz forms substantivized genitive expressions with the suffix *-NIKI*, which functions similarly to English "'s" in phrases like "John's" (meaning "John's possession" without specifying what is possessed). An example would be:

(5) *Мен Асаныкын көрдүм.* men Asan-NIKI-n kör-DI-m. I Asan-GEN.SUBST-ACC see-PST.DIR-1SG "I saw Asan's (possession)."

This construction, like the substantivized locative, contains information about both a possessor (Asan) and an understood possessed item, with morphological features (accusative case) applying to the possessed item rather than the possessor.

Kasieva et al. (2023) propose handling these forms similarly to substantivized *-DAGI* constructions, by splitting tokens to represent the two participants and their relationship. This creates consistency in the treatment of similar null-headed constructions while accurately representing the syntactic structure.

These approaches to null-headed clauses aim to capture the underlying syntactic relationships in a way that's both linguistically accurate and compatible with UD principles, though they inevitably involve some compromise between theoretical elegance and practical implementation. The solutions represent a thoughtful adaptation of UD guidelines to the specific challenges posed by Kyrgyz syntax, demonstrating how the framework can be extended to accommodate typologically diverse languages.

Inflection versus Derivation

The boundary between inflection and derivation represents a classic challenge in morphological analysis, with implications for syntactic annotation. This distinction becomes particularly relevant in Kyrgyz for several noun suffixes that traditional grammar doesn't classify as case suffixes but which function similarly in many contexts. These include *-LUU* (ornative, "having X"), *-sIz* (abessive, "without X"), *-DAy* (semblative, "like X"), and *-çA* (adverbial, "in the manner of X").

From a morphological perspective, these suffixes share characteristics with both inflectional and derivational morphology. Like case suffixes, they are highly productive, can attach to virtually any noun, and don't significantly change the lexical meaning of the base. However, like derivational morphemes, they can change the part of speech of the base and create forms that function differently syntactically from the base noun.

Consider the following example:

(6) *Адам катасыз болбос.* adam qata-sIz bol-BAs. person error-ABE be-NEG.FUT.IDF "A person won't be without errors."

Here, *катасыз* (qata-sIz, 'error-ABE') functions as a predicative adjective meaning "errorless" or "without errors". From an English perspective, this might suggest analyzing *-sIz* as a derivational suffix creating an adjective from a noun. However, from a Kyrgyz perspective, this suffix behaves more like a case marker, appearing in regular paradigmatic opposition with other case forms.

Kasieva et al. (2023) identify three possible approaches to analyzing these suffixes:

1. Treating them as deriving adjectives or adverbs from nouns (a derivational analysis);
2. Analyzing them as case marking, extending the standard inventory of cases (an inflectional analysis);
3. Treating them as cliticized postpositions (a syntactic analysis).

After considering the advantages and disadvantages of each approach, they opt primarily for the third solution – treating these morphemes as separate syntactic elements despite their integration into a single orthographic word with their host. This approach aligns with analyses of similar elements in related languages and preserves the productivity of these elements while distinguishing them from true case

suffixes. It does, however, introduce some complexity in tokenization and may seem counterintuitive from the perspective of native speakers, who perceive these as single words.

This solution contrasts with Benli (2023), who uses a mixed approach in the existing Kyrgyz UD treebank, sometimes treating these forms as derived words with varying POS tags and sometimes as inflected forms of the base noun. This inconsistency creates challenges for users of the treebank and complicates the development of computational tools.

The inflection/derivation boundary represents a fundamental challenge in morphological analysis, with no perfect solution for forms that exhibit characteristics of both. The approach proposed by Kasieva et al. (2023) represents a thoughtful compromise that prioritizes syntactic clarity while acknowledging the hybrid nature of these forms. It also creates greater consistency with analyses of similar constructions in related Turkic languages, facilitating cross-linguistic comparison and multilingual application development.

Proposed Solutions and Future Directions

Unified Approaches for Kyrgyz UD Annotation

Based on our analysis of the challenges in Kyrgyz UD annotation, we propose several unified approaches that balance linguistic accuracy with practical implementation and cross-linguistic consistency. These recommendations aim to create a solid foundation for future treebank development while addressing the specific needs of Kyrgyz syntactic representation.

For *copula tokenization,* we recommend treating non-past copula subject agreement morphemes as cliticized forms of the defective copula verb (*э*), with AUX as their POS tag. This approach provides the most linguistically accurate and consistent analysis across tense forms, capturing the parallel between non-past forms where the copula appears as a morpheme (e.g., *үйүңдөмүн* "I am at your house") and past forms where it appears as a separate word (e.g., *үйүңдө элем* "I was at your house"). While this approach does require subword tokenization, which adds complexity to processing, the linguistic benefits outweigh this disadvantage, particularly for a morphologically rich language where subword segmentation is often necessary anyway.

For *"small words"* like *да*, *эле*, *бар*, *жок*, and *керек*, we advocate for a function-based approach that prioritizes capturing their true syntactic behavior rather than forcing them into categories based on translation equivalents or surface form. For *да*, this means distinguishing between its discourse particle function (PART with discourse relation) and its intensifier/emphatic uses (ADV with advmod:emph relation). For words like *бар* and *жок*, despite their translation into English using verbal constructions, we recommend analyzing them as adjectives based on their syntactic distribution in Kyrgyz. This approach creates greater consistency within the treebank and with analyses of similar elements in related languages.

For *null-headed clauses,* we recommend using subtoken analysis for complex forms with *-DAGI* and *-NIKI* to provide the most complete representation of the linguistic structure. While this approach deviates somewhat from standard UD tokenization practices, it allows both the overt and understood elements to participate appropriately in syntactic relationships. For substantivized verbal adjectives, we recommend treating them as nominal dependents (obj, nsubj, etc.) rather than clausal ones (ccomp, csubj), reflecting their function in the sentence while maintaining information about their verbal origin. These approaches balance descriptive adequacy with practical implementation within UD constraints.

For the *inflection/derivation distinction,* we suggest treating productive "case-like" suffixes such as *-LUU*, *-sIz*, *-DAy*, and *-çA* as cliticized postpositions. This solution strikes a balance between capturing their productivity and distinguishing them from core case suffixes, while also creating consistency with analyses of similar elements in related languages. While this approach does increase tokenization complexity, it provides the most accurate representation of the syntactic relationships involved.

These recommendations aim to create a consistent annotation framework for Kyrgyz that can serve as a foundation for future treebank development. By addressing the specific challenges posed by Kyrgyz syntax, they contribute to the broader goal of making UD truly universal while respecting linguistic diversity.

Cross-Linguistic Consistency

One of the central goals of the Universal Dependencies project is to enable cross-linguistic comparison and multilingual application development. To this end, consistency in annotation practices

across Turkic language treebanks is essential, particularly for closely related languages like Kyrgyz and Kazakh where many constructions have direct parallels.

Washington et al. (2023) emphasized the need for unified approaches to annotation challenges shared across Turkic languages, such as pronominal locative expressions formed with *-ki* (equivalent to Kyrgyz *-GI*). Their comparative study of annotation approaches across treebanks highlighted significant inconsistencies even within the same language, creating challenges for cross-linguistic research and tool development.

For Kyrgyz, achieving cross-linguistic consistency requires careful consideration of how similar constructions are analyzed in related languages, particularly Kazakh. The solutions proposed in this paper aim to align with best practices in other Turkic language treebanks while respecting Kyrgyz-specific features. For example, our approach to copula tokenization parallels that used in Kazakh, while our treatment of "small words" draws on insights from both Kazakh and Turkish annotations.

At the same time, true cross-linguistic consistency doesn't mean identical treatment of superficially similar constructions. Languages may differ in how specific forms behave syntactically, and annotation should reflect these differences. For example, while the Turkish *-ki* and Kyrgyz *-GI* are etymologically related, their syntactic distribution differs in ways that may justify different annotation approaches.

As treebanks for more Turkic languages are developed, maintaining this balance between consistency and language-specific accuracy will be increasingly important. Regular communication between annotation teams working on different languages, shared documentation of challenging constructions, and periodic reviews of cross-linguistic consistency will all be valuable in achieving this goal.

Expanding Kyrgyz UD Resources

The current state of Kyrgyz UD resources represents just the beginning of what's needed for comprehensive language technology support. Future work should focus on several key areas:

1. *Expanding treebank size and diversity:* Current Kyrgyz treebanks remain small by international standards; increasing the number of annotated sentences and diversifying the text genres will improve resource robustness. A minimum target of 10,000 tokens would provide a more solid foundation for parser training, while 50,000+ tokens would enable development of more accurate parsing models. This expansion should be guided by corpus linguistics principles to ensure representative coverage of linguistic phenomena.

2. *Domain coverage:* Including scientific texts, spoken dialogues, social media content, and other genres will ensure broader coverage of linguistic phenomena and better support for real-world applications. Each domain introduces unique syntactic patterns and vocabulary; for example, social media text often features code-switching, colloquialisms, and non-standard syntax that aren't represented in formal written genres. A truly comprehensive treebank should sample across these registers.

3. *Enhanced documentation:* Detailed documentation of annotation decisions and language-specific guidelines will facilitate consistent annotation practices and make resources more accessible to new researchers. This should include explicit discussion of challenging constructions like those analyzed in this paper, with clear examples and justifications for annotation choices. Ideally, this documentation would be integrated with the treebank itself, allowing users to easily understand the rationale behind specific annotation decisions.

4. *Tool development:* Creating accurate dependency parsers for Kyrgyz based on expanded treebanks will enable larger-scale processing of Kyrgyz texts and support applications like machine translation and information extraction. Initial parser development could use existing multilingual models like UDPipe or Stanza, fine-tuned on Kyrgyz data, while more sophisticated models could be developed as data availability increases. Evaluation should consider both standard parsing metrics and performance on the specific challenging constructions identified in this paper.

5. *Parallel resources:* Developing parallel treebanks with other languages, particularly other Turkic languages, will support contrastive studies and multilingual applications. A Kyrgyz-Kazakh parallel treebank would be particularly valuable given the close relationship between these languages, while Kyrgyz-English resources would support machine translation development. These parallel resources should maintain consistent annotation across languages while respecting language-specific structures.

These efforts will require sustained collaboration between linguists, computer scientists, and Kyrgyz language specialists. Institutional support from universities, research organizations, and language technology initiatives will be essential for making significant progress. While developing these resources represents a substantial investment, the potential benefits for Kyrgyz language technology, education, and cultural preservation make it worthwhile.

The expansion of Kyrgyz UD resources should be viewed not as an isolated effort but as part of a broader movement to make language technology more inclusive and representative of linguistic diversity. By addressing the specific challenges of Kyrgyz annotation within the UD framework, we contribute to the development of truly universal approaches to computational linguistics that can accommodate the full range of human languages.

Conclusion

This paper has examined the current state and future prospects of Universal Dependencies resources for the Kyrgyz language, focusing on the specific challenges posed by Kyrgyz morphosyntactic structures within the UD framework. Despite being a national language with millions of speakers, Kyrgyz remains underresourced in computational linguistics, with limited treebank data and parsing tools compared to languages with similar speaker populations.

The analysis of key annotation challenges – copula tokenization, “small words”, null-headed clauses, and the inflection/derivation distinction – highlights areas where standard UD guidelines require thoughtful adaptation to accommodate Kyrgyz linguistic structures. The solutions proposed by researchers working on Kyrgyz UD treebanks represent creative compromises between linguistic accuracy and cross-linguistic consistency, demonstrating how the UD framework can be extended to typologically diverse languages.

These challenges are not merely technical issues but reflect fundamental questions about linguistic representation and the balance between language-specific accuracy and cross-linguistic comparability. By addressing them in a principled way, we contribute not only to Kyrgyz computational linguistics but also to the broader development of the UD framework as a truly universal tool for syntactic annotation.

The recommendations in this paper aim to provide a solid foundation for future Kyrgyz UD development, creating consistency in annotation practices while respecting the unique features of Kyrgyz syntax. By proposing unified approaches to challenging constructions, we hope to facilitate the expansion of Kyrgyz treebanks and the development of more accurate parsing tools for this important but underresourced language.

As work progresses on Kyrgyz UD resources, continued collaboration between linguists specializing in Kyrgyz and those working on other Turkic languages will be essential. This cooperation will ensure that the resulting resources are both linguistically accurate and maximally useful for computational applications, while also contributing to the broader goal of consistent annotation across related languages.

The development of robust UD resources for Kyrgyz will have significant practical implications beyond linguistic research. It will enable the creation of parsing tools, machine translation systems, and other NLP applications that can benefit Kyrgyz speakers in education, information access, and digital communication. In an increasingly digital world, access to language technology has become an important aspect of educational and economic opportunity; by expanding Kyrgyz language resources, we contribute to addressing digital linguistic inequality.

Furthermore, consistent annotation practices across Turkic languages will enhance the value of these resources for typological studies and multilingual processing. By developing approaches that work not just for Kyrgyz but potentially for all Turkic languages, we contribute to a more comprehensive understanding of this important language family and its computational representation.

In conclusion, while significant challenges remain in the development of comprehensive UD resources for Kyrgyz, the progress made so far demonstrates the feasibility of adapting the UD framework to this typologically distinct language. By continuing to build on this foundation with expanded treebanks, improved annotation guidelines, and more sophisticated computational tools, we can help ensure that Kyrgyz speakers benefit fully from advances in language technology while also enriching our understanding of linguistic diversity within universal frameworks.

Acknowledgments

We gratefully acknowledge the foundational work of researchers at Kyrgyz-Turkish Manas University, Swarthmore College, Saarland University, and other institutions in developing initial UD resources for Kyrgyz. Particular thanks go to Aida Kasieva, Gulnura Dzhumalieva, Anna Thompson, Murat Jumashev, Bermet Chontaeva, and Jonathan Washington for their pioneering work on Kyrgyz syntactic annotation within the UD framework. Their meticulous attention to the complexities of Kyrgyz syntax has laid the groundwork for all subsequent work in this area.

We also recognize the valuable contributions of the annotators and students who have participated in creating these resources and helped advance our understanding of Kyrgyz syntax within the UD framework. Their patience and careful attention to detail in working through complex annotation decisions has been essential to the development of high-quality resources.

The Universal Dependencies community as a whole deserves recognition for creating and maintaining a framework that can accommodate the diversity of human languages while facilitating cross-linguistic comparison. The openness of this community to discussing language-specific challenges and extending the framework where necessary has been crucial for its application to typologically diverse languages like Kyrgyz.

References

Aili, M., Mushajiang, W., Yibulayin, T., Liu, K.A. (2018) Universal dependencies for Uyghur. 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). P. 44-50. (in English)

Benli, İ. (2023) UD_Kyrgyz-KTMU: Universal Dependency treebank for Kyrgyz. GitHub repository: https://github.com/UniversalDependencies/UD_Kyrgyz-KTMU (in English)

Çöltekin, Ç., Doğruöz, A., Çetinoğlu, Ö. (2022) Resources for Turkish natural language processing: A critical survey. Language Resources and Evaluation. (in English)

Dzhumalieva, G.K., Kasieva, A.A., Musazhanova, S.J. (2023) Adaptacija terminov vjeb-projekta universal`nye zavisimosti na kyrgyzskij jazyk. Bulleten` of KRSU. 23(6): 71-75. [Dzhumalieva, G.K., Kasieva, A.A., Musazhanova, S.J. (2023) Adaptation of Web Project Terms for Universal Dependencies in the Kyrgyz Language. Bulletin of KRSU. 23(6): 71-75]. http://doi.org/10.36979/1694-500X-2023-23-6-71-75 (in Russian)

Kasieva, A., Knappen, J., Fischer, S., Teich, E. (2020) A new Kyrgyz corpus: sampling, compilation, annotation. Poster presented at: 42. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft. Hamburg, Germany. (in English)

Kasieva, A., Dzhumalieva, G., Thompson, A., Jumashev, M., Chontaeva, B., Washington, J. (2023) Issues of Kyrgyz syntactic annotation within the Universal Dependencies framework. In Proceedings of the XI International Conference on Computer Processing of Turkic Languages (TurkLang 2023). (in English)

Kornai, A. (2013) Digital Language Death. PLoS ONE 8(10): e77056. https://doi.org/10.1371/journal.pone.0077056 (in English)

Makazhanov, A., Sultangazina, A., Makhambetov, O., Yessenbayev, Z. (2015) 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). P. 338-350. (in English)

Merzhevich, T., Ferraz Gerardi, F. (2022) Introducing YakuToolkit. Yakut treebank and morphological analyzer. In Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages. P. 185-188. (in English)

Musazhanova, S.J., Kasieva, A.A., Dzhumalieva, G.K. (2023) Sintaksicheskaja annotacija kyrgyzskogo jazyka na osnove novossozdannogo korpusa. Vestnik Issyk-Kul`skogo universiteta. 54: 140-148. [Musazhanova, S.J., Kasieva, A.A., Dzhumalieva, G.K. (2023) Syntactic Annotation of the Newly-Created Kyrgyz Corpus. Bulletin of the Issyk-Kul University. 54: 140-148.] (in Russian)

Nivre, J., de Marneffe, M.C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., Zeman, D. (2016) Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of LREC. P. 1659-1666. (in English)

Sulubacak, U., Gokirmak, M., Tyers, F., Çöltekin, Ç., Nivre, J., Eryiğit, G. (2016) Universal Dependencies for Turkish. In Proceedings of COLING. The 26th International Conference on Computational Linguistics: Technical Papers. P. 3444-3454. (in English)

Taguchi, C. (2022) UD Tatar-NMCTT: Universal Dependency corpus for Tatar. GitHub repository: https://github.com/UniversalDependencies/UD_Tatar-NMCTT. (in English)

Thompson, A. (2021) Syntactic Parallelism and Structure in Kyrgyz Proverbs. Bachelor's thesis. Bryn Mawr College, Pennsylvania. (in English)

Tyers, F., Washington, J. (2015) 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). P. 276-289. (in English)

Tyers, F., Washington, J., Çöltekin, Ç., Makazhanov, A. (2017) An assessment of Universal Dependency annotation

guidelines for Turkic languages. In Proceedings of the Fifth International Conference on Turkic Language Processing (TurkLang). P. 276-297. (in English)

Tyers, F., Sheyanova, M., Washington, J. (2018) UD Annotatrix: An annotation tool for Universal Dependencies. In Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT). P. 10-17. (in English)

Washington, J.N., Ipasov, M., Tyers, F.M. (2012) A finite-state morphological transducer for Kyrgyz. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). P. 934-940. (in English)

Washington, J., Tyers, F., Salimzianov, I. (2022) Non-finite verb forms in Turkic exhibit syncretism, not multifunctionality. Folia Linguistica 56(3): 693-742. https://doi.org/10.1515/flin-2022-2045 (in English)

Washington, J., Çöltekin, Ç., Akkurt, F., Chontayeva, B. Eslami, S., Dzhumaliyeva, G., Kasiyeva, A., Kuzgun, A., Marşan, B., Taguchi, C. (2023) Strategies for the Annotation of Pronominalised Locatives in Turkic Universal Dependency Treebanks. ArXiv preprint. (in English)

Әдебиеттер

Джумалиева Г.К., Касиева А.А., Мусажанова С.Дж. Адаптация терминов веб-проекта универсальные зависимости на кыргызский язык // Вестник КРСУ. – 2023. – 23(6): 71-75. http://doi.org/10.36979/1694-500X-2023-23-6-71-75

Мусажанова С.Ж., Касиева А.А., Джумалиева Г.К. Синтаксическая аннотация кыргызского языка на основе вновь новосозданного корпуса // Вестник Иссык-Кульского университета. – 2023. – 54: 140-148.

Aili, M., Mushajiang, W., Yibulayin, T., Liu, K.A. (2018) Universal dependencies for Uyghur. 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). P. 44-50. (in English)

Benli, İ. (2023) UD_Kyrgyz-KTMU: Universal Dependency treebank for Kyrgyz. GitHub repository: https://github.com/UniversalDependencies/UD_Kyrgyz-KTMU (in English)

Çöltekin, Ç., Doğruöz, A., Çetinoğlu, Ö. (2022) Resources for Turkish natural language processing: A critical survey. Language Resources and Evaluation. (in English)

Kasieva, A., Knappen, J., Fischer, S., Teich, E. (2020) A new Kyrgyz corpus: sampling, compilation, annotation. Poster presented at: 42. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft. Hamburg, Germany. (in English)

Kasieva, A., Dzhumalieva, G., Thompson, A., Jumashev, M., Chontaeva, B., Washington, J. (2023) Issues of Kyrgyz syntactic annotation within the Universal Dependencies framework. In Proceedings of the XI International Conference on Computer Processing of Turkic Languages (TurkLang 2023). (in English)

Kornai, A. (2013) Digital Language Death. PLoS ONE 8(10): e77056. https://doi.org/10.1371/journal.pone.0077056 (in English)

Makazhanov, A., Sultangazina, A., Makhambetov, O., Yessenbayev, Z. (2015) 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). P. 338-350. (in English)

Merzhevich, T., Ferraz Gerardi, F. (2022) Introducing YakuToolkit. Yakut treebank and morphological analyzer. In Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages. P. 185-188. (in English)

Nivre, J., de Marneffe, M.C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., Zeman, D. (2016) Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of LREC. P. 1659-1666. (in English)

Sulubacak, U., Gokirmak, M., Tyers, F., Çöltekin, Ç., Nivre, J., Eryiğit, G. (2016) Universal Dependencies for Turkish. In Proceedings of COLING. The 26th International Conference on Computational Linguistics: Technical Papers. P. 3444-3454. (in English)

Taguchi, C. (2022) UD Tatar-NMCTT: Universal Dependency corpus for Tatar. GitHub repository: https://github.com/UniversalDependencies/UD_Tatar-NMCTT. (in English)

Thompson, A. (2021) Syntactic Parallelism and Structure in Kyrgyz Proverbs. Bachelor's thesis. Bryn Mawr College, Pennsylvania. (in English)

Tyers, F., Washington, J. (2015) 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). P. 276-289. (in English)

Tyers, F., Washington, J., Çöltekin, Ç., Makazhanov, A. (2017) An assessment of Universal Dependency annotation guidelines for Turkic languages. In Proceedings of the Fifth International Conference on Turkic Language Processing (TurkLang). P. 276-297. (in English)

Tyers, F., Sheyanova, M., Washington, J. (2018) UD Annotatrix: An annotation tool for Universal Dependencies. In Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT). P. 10-17. (in English)

Washington, J.N., Ipasov, M., Tyers, F.M. (2012) A finite-state morphological transducer for Kyrgyz. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). P. 934-940. (in English)

Washington, J., Tyers, F., Salimzianov, I. (2022) Non-finite verb forms in Turkic exhibit syncretism, not multifunctionality. Folia Linguistica 56(3): 693-742. https://doi.org/10.1515/flin-2022-2045 (in English)

Washington, J., Çöltekin, Ç., Akkurt, F., Chontayeva, B. Eslami, S., Dzhumaliyeva, G., Kasiyeva, A., Kuzgun, A., Marşan, B., Taguchi, C. (2023) Strategies for the Annotation of Pronominalised Locatives in Turkic Universal Dependency

Treebanks. ArXiv preprint. (in English)

Information about the article / Мақала туралы ақпарат / Информация о статье
Entered the editorial office / Редакцияға түсті / Поступила в редакцию: 16.03.2025.
Accepted for publication / Жариялауға қабылданды / Принята к публикации: 25.06.2025.