From variation to harmonization: the UD Turkic Group initiative

Soudabeh Eslami[1], Furkan Akkurt[2], Bermet Chontaeva[1], Çağrı Çöltekin[1], Nikolett Mus[3], Jonathan Washington[4],
[1]University of Tübingen, [2]Boğaziçi University [3]ELTE Research Centre for Linguistics, [4]Swarthmore College
{soudabeh.eslami, bermet.chontaeva}@student.uni-tuebingen.de
furkan.akkurt@bogazici.edu.tr
cagri.coeltekin@uni-tuebingen.de
mus.nikolett@nytud.elte.hu
jonathan.washington@swarthmore.edu

Relevant UniDive working groups: WG1 (Corpus annotation), WG4 (Quantifying and promoting diversity)
Track: Work-in-progress

1 Introduction

This abstract summarizes the activities of the UD Turkic Group, established in September 2023 within the framework of the UniDive COST Action (CA21167) and operating primarily within WG1. The group aims to harmonize annotations across Universal Dependencies (UD) treebanks for Turkic languages and to develop guidelines for future initiatives.

As of UD v2.17, 26 treebanks covering 12 Turkic languages or varieties exhibit considerable cross-treebank variation. These discrepancies arise from both unresolved issues in linguistic analysis and independent design choices across annotation efforts. Although Turkic languages share typologically and genetically grounded features, these are not always analysed uniformly, due to differing linguistic traditions and gaps in language-specific descriptions.

To address this, the group conducts collaborative cross-linguistic analysis with the goal of reconciling divergent practices and formulating recommendations for consistent and interoperable annotation across Turkic UD treebanks.

2 Background

The Turkic language family spans a vast area from southeastern Europe to northeastern Asia, written in Latin, Cyrillic, Arabic, and historical scripts. In UD, Turkic languages are represented across the Oghuz (Turkish, Ottoman Turkish, Azerbaijani), Karluk (Uzbek, Uyghur), Kipchak (Kazakh, Kyrgyz, Tatar), and Siberian (Sakha) branches, alongside Old Turkish and two code-switching corpora.

Coverage is highly uneven. Turkish alone accounts for 11 treebanks totaling over 700,000 tokens across diverse genres including news, web text, grammar examples, dictionary entries, tourism reviews, and spoken data. Most other languages are represented by only one or two treebanks, often with fewer than 10,000 tokens, and several consist of only a few hundred sentences. The larger treebanks tend to be well-established and regularly updated, while many of the smaller ones represent initial annotation efforts.

Known divergences include differences in copula tokenization and lemmatization – where even the major Turkish treebanks follow mutually inconsistent practices – the morphological analysis of auxiliaries, the treatment of pronominal locative constructions, and varying morphological feature inventories across teams and languages.

3 Methodology

Since its formation, the UD Turkic Group has coordinated its efforts through regular online meetings, during which subgroups were established to address specific tasks. These subgroups focused on mapping annotation inconsistencies, surveying relevant linguistic literature, and formulating linguistically oriented research questions to guide harmonization. The process was structured around two in-person workshops. The inaugural UD Turkic Workshop (Istanbul, September 2023), co-located with the UniDive 2nd WG3 meeting, convened treebank developers and Turkic linguists to review existing treebanks, identify inconsistencies, and set a shared research agenda. The UD Turkic Workshop (Ljubljana, August 2025), held alongside SyntaxFest 2025, provided a forum for presenting ongoing annotation work and parallel treebank projects, facilitating discussion of emerging challenges and consolidation of optimal practices. Together, the online and in-person activities formed a systematic methodology, enabling cross-linguistic comparison, alignment of UD annotations, and the development of collaborative standards for Turkic

treebanks.

4 Research topics and publications

The group has concentrated on a number of recurrent annotation problems underlying cross-treebank inconsistencies in Turkic UD, focusing on cases where shared structural features receive divergent analyses. Phenomena occurring at the interface of morphology and syntax appear to pose particular annotation challenges. This work has led to a series of conference presentations and ongoing publications.

One such issue concerns pronominalized locatives, i.e. constructions in which case-marked pronouns function as locative expressions. A cross-treebank survey revealed divergent annotation strategies, which were systematically reviewed by Washington et al. (2024) (MWE-UD 2024, LREC-COLING, Turin), leading to the proposal of a more consistent treatment.

Another area of variation involves the tokenization of copula constructions, which in many Turkic languages are realized as clitics or suffixes. Based on a comparative analysis, Coltekin et al. (2026) (SIGTURK 2026, EACL, Rabat) proposes a unified tokenization strategy grounded in cross-linguistic evidence.

Closely related issues arise in the analysis of non-verbal negation. A recent study (accepted at UDW 2026) examining Azerbaijani, Kyrgyz, and Turkish demonstrates that the negative element in such constructions cannot be consistently analyzed as an auxiliary or copula. Instead, it is treated as `ADV` with the dependency relation `advmod:neg`, based on its distributional and morphological properties.

In addition to these targeted studies, the group has developed parallel UD treebanks for multiple Turkic languages, currently available for Azerbaijani, Kyrgyz, Turkish, and Uzbek (Akhundjanova et al., 2025), with further languages in progress. Aligned at the sentence level, these resources enable direct cross-linguistic comparison and provide a testing ground for harmonized annotation strategies. The resource was presented at UDW 2025 (SyntaxFest, Ljubljana).

5 Impact on UD Treebanks

A central question for the group is whether treebanks outside the group's direct control have adopted its proposed analyses. While the TueCL treebanks (maintained by group members) naturally reflect current recommendations, uptake in independently maintained treebanks is what measures the group's broader impact. We examined several established treebanks — including Turkish-BOUN, Turkish-IMST, Ottoman Turkish-BOUN, Kyrgyz-KTMU, and Tatar-NMCTT — checking their annotation practices and commit histories against the three main proposals.

For pronominalized locatives, the paper proposes segmenting pronominal *-ki* forms into multiword tokens and assigning the *-ki* subword a nominal POS tag. While most treebanks do segment these forms, they assign *-ki* entirely different tags: BOUN uses `PART` with `dep:der`, IMST and PUD use `ADP` with `case`, and GB uses a mix of `ADP` and `SCONJ`. No non-TueCL treebank uses the proposed nominal analysis.

For copula tokenization, Turkish-BOUN and Turkish-IMST have long-standing but mutually inconsistent tokenization practices (e.g., different copula lemmas), and neither has adopted the unified policy.

For non-verbal negation, the proposed `advmod:neg` relation has zero occurrences across all 26 Turkic treebanks; all independently maintained treebanks analyze the negation element as `AUX` with `aux` or `cop`.

These findings indicate that the group's proposals have not yet been adopted beyond the treebanks maintained by its own members. This highlights the need for more active dissemination — such as proposing changes directly to treebank maintainers and coordinating updates through the UD release cycle — and motivates the group's continued engagement through workshops and publications.

6 Future plans

Future work will focus on extending the parallel treebanks — currently published for four languages, with five more in progress (Crimean Tatar, Karakalpak, Kumyk, Sakha, and Tatar) and further expansion planned to Kazakh and Uyghur — and on addressing additional annotation challenges such as converb constructions, multi-word expressions, and postpositional phrases. A key objective is to promote broader adoption by coordinating directly with maintainers of independently developed treebanks and integrating recommendations into the UD release cycle.

Acknowledgements

This work is supported by COST Action CA21167 — Universality, diversity and idiosyncrasy in language technology (UniDive).

References

Arofat Akhundjanova, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, and Cagri Coltekin. 2025. Parallel Universal Dependencies Treebanks for Turkic Languages. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 129–136, Ljubljana, Slovenia. Association for Computational Linguistics.

Cagri Coltekin, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, Sardana Ivanova, Gulnura Dzhumalieva, Aida Kasieva, Nikolett Mus, and Jonathan Washington. 2026. Tokenisation of Turkic Copula Constructions in Universal Dependencies. In *Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026)*, pages 172–178, Rabat, Morocco. Association for Computational Linguistics.

Jonathan Washington, Çağrı Çöltekin, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, Gulnura Jumalieva, Aida Kasieva, Aslı Kuzgun, Büşra Marşan, and Chihiro Taguchi. 2024. Strategies for the Annotation of Pronominalised Locatives in Turkic Universal Dependency Treebanks. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 207–219, Torino, Italia. ELRA and ICCL.