Parallel Universal Dependencies Treebanks for Turkic Languages

Arofat Akhundjanova[1], Furkan Akkurt[2],
Bermet Chontaeva[3], Soudabeh Eslami[3], Çağrı Çöltekin[3]
[1]Independent Researcher, [2]Boğaziçi University [3]University of Tübingen,
arofat.akhundjanova@gmail.com furkan.akkurt@bogazici.edu.tr
{bermet.chontaeva, soudabeh.eslami}@student.uni-tuebingen.de
cagri.coeltekin@uni-tuebingen.de

Abstract

We introduce a set of fully aligned and manually annotated parallel Universal Dependencies (UD) treebanks for four Turkic languages: Azerbaijani, Kyrgyz, Turkish, and Uzbek. The treebanks were annotated with close collaboration between annotators to ensure the harmonized annotations across all languages. These resources currently consist of 148 strategically selected sentences that illustrate typologically significant morphosyntactic phenomena across these related languages. These parallel treebanks enable systematic comparative studies of Turkic syntax and may be instrumental in cross-lingual NLP applications. All treebanks are available as part of UD v2.16.

1 Introduction

The Universal Dependencies (UD) framework has emerged as the leading standard for cross-linguistically consistent grammatical annotation, covering over 150 languages across more than 300 treebanks (Nivre et al., 2020). However, the coverage and depth of resources for Turkic languages are limited and vary considerably in treebank size and availability across the language family. More critically, the absence of parallel treebanks for these languages limits systematic cross-linguistic comparisons, despite their shared typological characteristics and historical ties. To the best of our knowledge, there are only two parallel treebanking efforts that include a Turkic language (Turkish in both): (1) the PUD treebank, created for the CoNLL 2017 Shared Task on Multilingual Parsing (Zeman et al., 2017), (2) the Atis treebank (Cesur et al., 2024). The PUD treebank consists of 1000 manually translated sentences in 22 languages, including Turkish. The Atis treebank contains a translation of the English ATIS (Airline Travel Information System) corpus to Turkish.

As of UD version 2.16, there are 24 treebanks for 11 Turkic languages (including historical and code-switching varieties, see Table 1 in Appendix A for an overview). As pointed out by multiple earlier studies, annotation inconsistencies are common in Turkic treebanks (Türk et al., 2019; Çöltekin et al., 2023). Some of these issues were addressed in the recent work, such as pronominalized locatives (Washington et al., 2024) and postverbial constructions (Akhundjanova, 2025). However, there remains a notable gap in cross-linguistic research based on parallel corpora for Turkic languages.

This paper addresses this gap by introducing the first fully aligned parallel UD treebanks for four Turkic languages: Azerbaijani, Kyrgyz, Turkish, and Uzbek. These languages represent distinct branches of the Turkic family—Oghuz (Azerbaijani, Turkish), Kipchak (Kyrgyz), and Karluk (Uzbek)—allowing for meaningful cross-branch comparisons. Even though most treebanks introduced in this work were annotated by a single annotator (one of the authors), the treebanks are cross-checked by others to ensure the consistency of annotations as well as their quality.

Our approach combines carefully constructed grammatical examples with linguistic expertise to create resources that highlight both the shared typological features and unique syntactic traits of each language.

The treebanks are available as part of Universal Dependencies v2.16 and support applications in cross-lingual parsing, comparative research, and language education.

2 Languages and Corpora

2.1 Overview of Selected Languages

Turkic languages are agglutinative, head-final, and typically pro-drop (Johanson, 2021). The four languages covered in our study represent three

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 129–136

Language	UD Treebanks	Size	Parallel?
Azerbaijani	TueCL	Small	Yes
Kazakh	KTB	Small	No
Kyrgyz	TueCL, KTMU	Medium	Yes
Tatar	NMCTT	Small	No
Turkish	Kenet, Penn, Tourism, Atis, GB, FrameNet, IMST, BOUN, PUD, DUDU, Tonqq	Large	Yes (PUD, Atis)
Uyghur	UDT	Medium	No
Uzbek	UDT	Small	No
Yakut	YKTDT	Small	No

Table 1: Status of UD treebanks for Turkic languages as of version 2.15. Treebank sizes are classified based on token counts: small (under 20K tokens), medium (20K–100K tokens), and large (over 100K tokens). The "Parallel?" column indicates whether the language has a treebank which is part of a cross-linguistic parallel corpus. The parallel treebanks for Azerbaijani and Kyrgyz are also part of the current study.

distinct branches of the Turkic family: Azerbaijani and Turkish from the Oghuz branch, Kyrgyz from the Kipchak branch, and Uzbek from the Karluk branch (Johanson, 2022). While they share core typological features, notable differences in morphology and syntax make them well-suited for comparative analysis.

Common challenges in Turkic UD treebanks arise from several language-specific features. Agglutinative morphology leads to long, complex word forms that complicate tokenization and annotation processes. Syntactic analysis is further challenged by complex verbal constructions, such as serial verbs and auxiliary chains. Moreover, lower-resource languages like Azerbaijani, Kyrgyz, and Uzbek lack robust automatic parsing tools, limiting corpus development and comprehensive analysis.

2.2 Source Data

Our parallel treebanks are based on a curated collection of 148 sentences, compiled from multiple sources with linguistic annotation in mind. Specifically, we draw from the Cairo corpus (20 sentences),[1] the UDTW23 corpus (20 sentences),[2] and 97 additional examples illustrating grammatical constructions of interest. We added 9 sentences to the Cairo corpus to capture alternative annotations for pronominal subject omission. Although they appear duplicated,

[1] https://github.com/UniversalDependencies/cairo

[2] https://github.com/ud-turkic/udtw23

Statistic	AZ	KY	TR	UZ
Tokens	912	1048	904	940
Avg. sent. length	6.2	7.1	6.1	6.4
POS tags	15	16	14	15
Dependencies	34	38	37	33
Avg. dep. length	2.3	2.4	2.3	2.4

Table 2: Basic statistics for the parallel treebanks. AZ: Azerbaijani, KY: Kyrgyz, TR: Turkish, UZ: Uzbek.

each has a unique sentence ID and reflects a distinct annotation—with and without pronouns. Two new sentences were added to the original UDTW23 corpus for additional analysis of language-specific issues in Kyrgyz. Most of the source sentences originate in Turkish and were manually translated into Azerbaijani, Kyrgyz, and Uzbek, preserving semantic and syntactic alignment across languages. The tokenized data in the Turkish treebank amounts to 904 tokens. Additionally, English translations are provided as metadata to facilitate cross-lingual comparisons. Table 2 shows basic statistics for the parallel treebanks.

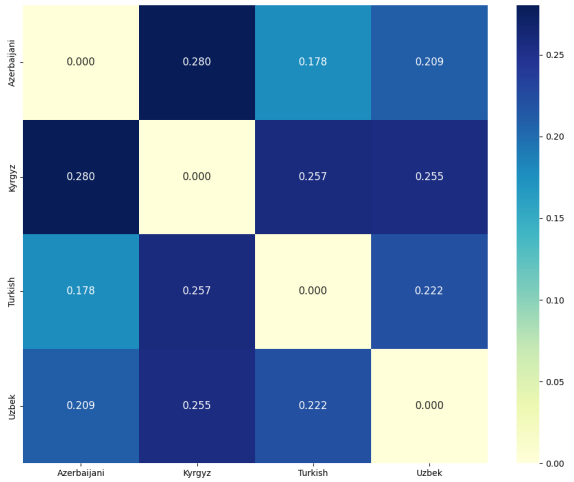
In translating the data, we prioritized maintaining structural alignment while ensuring idiomatic and grammatical correctness in each target language. This allows for consistent parallelism that supports both syntactic and semantic cross-linguistic analyses.

Sentences were selected based on their value for (cross)linguistic analysis, with a focus on morphosyntactically rich or typologically marked constructions. These include pro-drop sentences, auxiliary chains, postverbal structures, and non-canonical word orders—phenomena that are both theoretically significant and empirically challenging for dependency annotation and parser development.

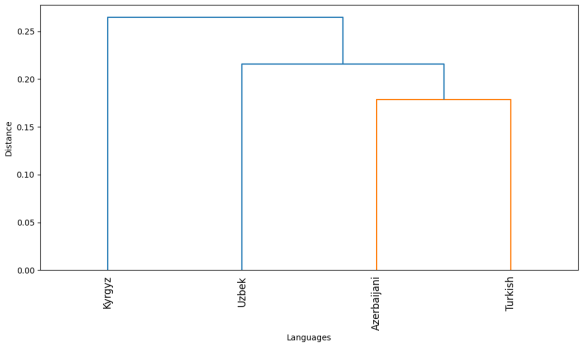
The treebanks are encoded in Latin script for Azerbaijani, Turkish, and Uzbek. The Kyrgyz corpus is presented in its native Cyrillic script, with transliteration and interlinear glosses provided in the metadata (see Appendix B for an annotation sample).

3 Treebank Construction

This section outlines our treebank development process, including the general annotation methodology and language-specific approaches for each treebank. We also provide a quantitative analysis of the parallel corpora to highlight typological similarities and distinctive features of



(a) Normalized edit distances based on POS sequences.



(b) The dendrogram for language clustering, showing structural similarities among the languages.

Figure 1: Normalized edit distances between languages and language clustering based on these distances.

these languages. Additionally, we discuss some annotation issues and describe our approach to resolving them.

3.1 Annotation Workflow

Our approach integrates automated processing with manual annotation and revision. When feasible, processing tasks were initially automated and then refined by native speakers with relevant expertise. Each treebank underwent thorough manual verification and correction, in line with the UD guidelines.[3] Ambiguous cases were addressed through collaborative discussions with linguists and UD experts.

The development of the four treebanks followed different paths based on existing resources and language-specific requirements. The Azerbaijani and Kyrgyz treebanks, both named *TueCL*, were created prior to the current project (Eslami and Çöltekin, 2024; Chontaeva and Çöltekin, 2024). They were enriched with additional grammar examples and morphological features as part of this project. Turkish-TueCL and Uzbek-TueCL, in contrast, were developed from scratch with distinct approaches.

For Turkish, we first followed two different strategies independently, (1) fully manual annotation and (2) automatic annotation followed by manual correction, after which we manually merged the annotations. For automatic annotations, we used the large language model Claude 3.5 Sonnet (Anthropic, 2025).

Regarding Uzbek, tokenization was automated with the NLTK,[4] but all other annotation layers were performed manually.

3.2 Quantitative Analysis

Using parallel treebanks, we performed a quantitative analysis to highlight structural and typological features of the target languages. It is important to acknowledge that the use of constructed sentences may limit the generalizability of our findings to more natural language use.

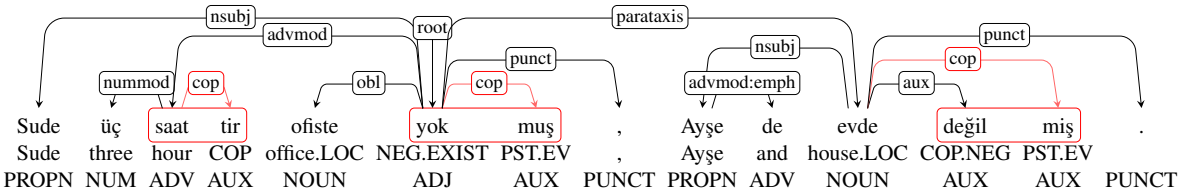
Figure 1a presents the normalized edit distances between languages based on POS sequences, while Figure 1b shows the resulting language clustering derived from these distances. The results show that Azerbaijani and Turkish cluster closely with the shortest edit distance, while Kyrgyz shows the longest distances, particularly from Azerbaijani, reflecting less POS alignment and greater structural divergence. Overall, these findings confirm that typological relationships between languages are reflected in our annotations.

Among the languages studied, Turkish exhibits the shortest dependency lengths (up to 15 tokens), whereas Uzbek displays the longest (up to 21 tokens). This variation may stem from differences in translation styles and structural preferences.

In terms of word order, all languages predominantly adhere to a subject-object-verb (SOV) structure. Relative clauses consistently precede head nouns, conforming to the typical

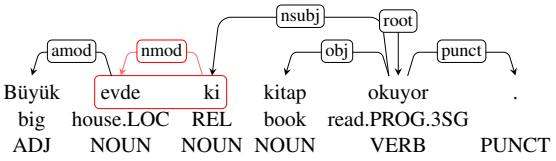
[3] https://universaldependencies.org/guidelines

[4] http://www.nltk.org/api/nltk.tokenize.html



‘Sude was not at the office for three hours and Ayşe was not at home.’

Figure 2: Example annotation of several copular constructions in Turkish. Copular elements are treated as separate subtokens from the syntactic head and are highlighted within red frames.



‘The one in the big house is reading (a) book.’

Figure 3: Example annotation of a pronominalized locative in Turkish with a *-ki* construction treated as a separate subtoken from the syntactic head and highlighted within a red frame.

head-final nature of Turkic syntax. In modifier constructions, Turkish often places determiners after adjectives, while the reverse is more common in the other languages, though both patterns are acceptable depending on stylistic variation and pragmatic factors.

3.3 Unique Features

Our cross-linguistic analysis revealed not only the common characteristics mentioned above but also unique features specific to individual languages. One such feature is the flexible positioning of the question particle *mi* in Turkish, which can shift the focus of a sentence depending on its placement. In contrast, the other languages in our study use question markers with a fixed position, consistently appearing at the end of the predicate.

Another feature related to question markers is the tendency in Azerbaijani to form questions without particles, both in written and conversational speech. However, such intonation-based questions are not widely used in the formal contexts of the other languages.

The use of posture and locational verbs, e.g., жат (‘jat’ to lie down), as an auxiliary to mark the progressive aspect is more characteristic and prominent in Kyrgyz. Uzbek also shows similar aspectual chaining with postverbal constructions, but such usage is usually limited to conversational speech. In contrast, in Turkish, such auxiliary constructions are not very productive and progressive aspect is primarily marked by inflectional morphology.

Another feature unique to Kyrgyz is the ability to form compound nouns without a possessive suffix—for example, алма даректар (alma daraktar, ‘apple trees’)—whereas other languages in this study require a possessive suffix on the head noun (e.g., *olma daraxtlari* in Uzbek).

3.4 Annotation Challenges

Annotating Turkic languages in the UD framework has some challenges discussed in earlier studies (Tyers et al., 2017), some of which have been under active discussion recently (Taguchi, 2022). Here, we briefly note two issues where we adopt the current consensus that resulted from these recent discussions. The first is the copular constructions in Turkic languages, where copula can be realized as an affix attached to a nominal or adjectival predicate, making syntactic and morphological analysis difficult, and may result in inconsistent analyses of similar constructions. We treat copular affixes as separate syntactic units with the AUX POS tag and attach them to the main predicate with the cop dependency relation to better reflect their syntactic behavior. Figure 2 presents an example analysis with multiple copular affixes.

Another of the issues is pronominalized locatives with the suffix *-ki* as discussed in Washington et al. (2024). We follow the analysis in that work and treat *-ki* constructions as separate subtokens, as shown in Figure 3. Albeit making its automatic annotation non-trivial, this approach preserves all the linguistic information in both the genitive and locative variants of this construction.

4 Conclusion and Future Work

We presented the first fully aligned parallel Universal Dependencies treebanks for four Turkic languages. This work fills a significant gap in resources for comparative studies of Turkic syntax and provides valuable data for cross-lingual NLP applications.

Our future work includes extending these treebanks with additional texts from various genres and adding more Turkic languages, particularly from less-represented branches of the family. We also plan to conduct more detailed analyses of the morphosyntactic phenomena highlighted in this initial study.

We invite the research community to collaborate on this ongoing project to improve the representation of Turkic languages in language resources and models.

Limitations

Our parallel treebanks have several limitations. First, the relatively small size (148 sentences) may limit their utility for certain applications, particularly data-hungry machine learning tasks. Second, the focus on constructed examples, while valuable for highlighting specific linguistic phenomena, may not fully represent natural language usage. Third, the current version primarily addresses written language and formal registers.

Despite these limitations, these parallel treebanks provide a valuable starting point for cross-linguistic Turkic language studies and demonstrate the feasibility of our annotation approach for future extensions.

Acknowledgments

We thank the Turkic UD working group for fruitful discussions of linguistic issues and annotation approaches. This work was supported by COST Action CA21167 – Universality, diversity and idiosyncrasy in language technology (UniDive).

References

Arofat Akhundjanova. 2025. Harmonizing annotation of Turkic postverbial constructions: A comparative study of UD treebanks. In *Proceedings of the 18th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 10–17, Abu Dhabi, UAE. Association for Computational Linguistics.

Arofat Akhundjanova and Luigi Talamo. 2025. Universal Dependencies treebank for Uzbek. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 1–6, Tallinn, Estonia. University of Tartu Library, Estonia.

Furkan Akkurt, Nursena Teker, Helin Binici, Ahmet Demir, and Konstantinos Sampanis. 2025. UD Turkish-English BUTR. https://github.com/UniversalDependencies/UD_Turkish_English-BUTR.

Anthropic. 2025. Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 2025.

Ibrahim Benli. 2020. UD Kyrgyz KTMU. https://github.com/UniversalDependencies/UD_Kyrgyz-KTMU.

Neslihan Cesur, Aslı Kuzgun, Mehmet Kose, and Olcay Taner Yıldız. 2024. Building annotated parallel corpora using the ATIS dataset: Two UD-style treebanks in English and Turkish. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC)@ LREC-COLING 2024*, pages 104–110.

Neslihan Cesur, Aslı Kuzgun, Olcay Taner Yıldız, Büşra Marşan, Neslihan Kara, Bilge Nas Arıcan, Merve Özçelik, and Deniz Baran Aslan. 2022. UD Turkish Penn. https://github.com/UniversalDependencies/UD_Turkish-Penn.

Özlem Çetinoğlu and Çağrı Çöltekin. 2022. Two languages, one treebank: Building a Turkish–German code-switching treebank and its challenges. *Language Resources and Evaluation*, pages 1–35.

Bermet Chontaeva and Çağrı Çöltekin. 2024. UD Kyrgyz TueCL. https://github.com/UniversalDependencies/UD_Kyrgyz-TueCL.

Çağrı Çöltekin. 2015. A grammar-book treebank of Turkish. In *Proceedings of the 14th Workshop on Treebanks and Linguistic Theories (TLT 14)*, pages 35–49.

Çağrı Çöltekin, A. Seza Doğruöz, and Özlem Çetinoğlu. 2023. Resources for Turkish natural language processing: A critical survey. *Language Resources and Evaluation*, 57(1):449–488.

Mehmet Oguz Derin and Takahiro Harada. 2021. Universal Dependencies for Old Turkish. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 129–141, Sofia, Bulgaria. Association for Computational Linguistics.

Marhaba Eli, Weinila Mushajiang, Tuergen Yibulayin, Kahaerjiang Abiderexiti, and Yan Liu. 2016. Universal dependencies for Uyghur. 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), pages 44–50, Osaka, Japan. The COLING 2016 Organizing Committee.

Soudabeh Eslami and Çağrı Çöltekin. 2024. UD Azerbaijani TueCL. https://github.com/UniversalDependencies/UD_Azerbaijani-TueCL.

Lars Johanson. 2021. The structure of Turkic. In Lars Johanson and Éva Á. Csató, editors, *The Turkic Languages*, pages 26–59. Routledge.

Lars Johanson. 2022. The history of Turkic. In Lars Johanson and Éva Á. Csató, editors, *The Turkic Languages*, pages 83–123. Routledge.

Aslı Kuzgun, Neslihan Cesur, Olcay Taner Yıldız, Oğuzhan Kuyrukçu, Büşra Marşan, Bilge Nas Arıcan, Neslihan Kara, Deniz Baran Aslan, Ezgi Sanıyar, and Cengiz Asmazoğlu. 2023. Turkish tourism treebank of Universal Dependencies 2.13. https://hdl.handle.net/11234/1-5287.

Aslı Kuzgun, Neslihan Cesur, Olcay Taner Yıldız, Oğuzhan Kuyrukçu, Arife Betül Yenice, Bilge Nas Arıcan, and Ezgi Sanıyar. 2022. UD Turkish Kenet. https://github.com/UniversalDependencies/UD_Turkish-Kenet.

Büşra Marşan, Neslihan Kara, Merve Özçelik, Bilge Nas Arıcan, Neslihan Cesur, Aslı Kuzgun, Ezgi Sanıyar, Oğuzhan Kuyrukçu, and Olcay Yıldız. 2021. Building the Turkish FrameNet. In *Proceedings of the 11th Global Wordnet Conference*, pages 118–125.

Tatiana Merzhevich and Fabrício Ferraz Gerardi. 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*, pages 185–188.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Şaziye Özateş, Tarık Tıraş, Efe Genç, and Esma Bilgin Tasdemir. 2024. Dependency annotation of Ottoman Turkish with multilingual BERT. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 188–196, St. Julians, Malta. Association for Computational Linguistics.

Umut Sulubacak, Memduh Gokirmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal Dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454, Osaka, Japan. The COLING 2016 Organizing Committee.

Chihiro Taguchi. 2022. Consistent grammatical annotation of Turkic languages for more universal Universal Dependencies. In *29th International Conference on Head-Driven Phrase Structure Grammar*.

Chihiro Taguchi, Sei Iwata, and Taro Watanabe. 2022. Universal Dependencies treebank for Tatar: Incorporating intra-word code-switching information. In *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, pages 95–104, Marseille, France. European Language Resources Association.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2019. Improving the annotations in the Turkish universal dependency treebank. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 108–115.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Gözde Berk, Seyyit Talha Bedir, Abdullatif Köksal, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2021. Resources for Turkish dependency parsing: Introducing the BOUN treebank and the BoAT annotation tool. *Preprint*, arXiv:2002.10416.

Francis Tyers, Jonathan Washington, Çağrı Çöltekin, and Aibek Makazhanov. 2017. An assessment of Universal Dependency annotation guidelines for Turkic languages. In *5th International Conference on Turkic Language Processing (TURKLANG 2017)*, pages 356–377.

Francis M. Tyers and Jonathan N. Washington. 2015. Towards a free/open-source universal-dependency treebank for Kazakh. In *3rd International Conference on Turkic Languages Processing (TurkLang 2015)*, pages 276–289.

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.

Enes Yılandiloğlu and Janine Siewert. 2025. DUDU: A treebank for Ottoman Turkish in UD style. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 74–79, Tallinn, Estonia. University of Tartu Library, Estonia.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, and 43 others. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

A Turkic UD Treebanks

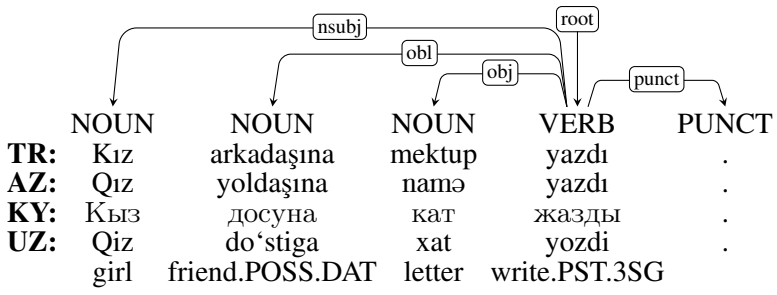
This appendix provides a comprehensive overview of all Turkic language treebanks available in Universal Dependencies version 2.16. These treebanks vary significantly in size, coverage, and domain, highlighting the uneven resource distribution across the Turkic language family. Table 3 presents key statistics for each treebank.

Treebank	Sentences	Tokens	Genre	Additional Information
Azerbaijani-TueCL (Eslami and Çöltekin, 2024)	148	912	grammar	Grammar examples
Kazakh-KTB (Tyers and Washington, 2015)	1078	10 536	news, fiction, wiki	Mixed sources
Kyrgyz-KTMU (Benli, 2020)	2480	23 654	news, fiction	Media texts
Kyrgyz-TueCL (Chontaeva and Çöltekin, 2024)	173	1250	grammar	Grammar examples
Tatar-NMCTT (Taguchi et al., 2022)	148	2280	news, non-fiction	Mixed sources
Turkish-Atis (Cesur et al., 2024)	5432	45 907	news, non-fiction	ATIS domain
Turkish-BOUN (Türk et al., 2021)	9761	125 212	news, non-fiction	Web sources
Turkish-FrameNet (Marşan et al., 2021)	2698	19 223	grammar	FrameNet annotations
Turkish-GB (Çöltekin, 2015)	2880	17 177	grammar	Grammar examples
Turkish-IMST (Sulubacak et al., 2016)	5635	58 096	news, non-fiction	METU-Sabancı corpus
Turkish-Kenet (Kuzgun et al., 2022)	18 687	178 658	grammar	Corpus of example sentences
Turkish-Penn (Cesur et al., 2022)	16 396	183 555	news, non-fiction	Turkish Penn Treebank
Turkish-PUD (Zeman et al., 2017)	1000	16 881	news, wiki	Parallel corpus
Turkish-Tourism (Kuzgun et al., 2023)	19 830	91 152	reviews	Hotel reviews
Turkish-German-SAGT (Çetinoğlu and Çöltekin, 2022)	2184	37 227	spoken	Code-switching corpus
Turkish-English-BUTR (Akkurt et al., 2025)	51	393	spoken	Code-switching corpus
Old Turkish-Tonqq (Derin and Harada, 2021)	20	158	non-fiction	Historical texts
Ottoman Turkish-BOUN (Özateş et al., 2024)	514	8834	fiction, non-fiction	Historical texts
Ottoman Turkish-DUDU (Yılandiloğlu and Siewert, 2025)	1064	10 012	bible, fiction, non-fiction, government, news	Mixed sources
Uyghur-UDT (Eli et al., 2016)	3456	40 236	fiction	Literary texts
Uzbek-UT (Akhundjanova and Talamo, 2025)	500	5850	news, fiction	Mixed sources
Yakut-YKTDT (Merzhevich and Gerardi, 2022)	299	1459	news, non-fiction	Media texts

Table 3: Basic statistics on Turkic UD treebanks available in version 2.16, excluding the Turkish and Uzbek treebanks introduced in this paper.

B Annotation sample of parallel sentences

sent_id = cairo-1
text[tr] = Kız arkadaşına mektup yazdı.
text[az] = Qız yoldaşına məktub yazdı.
text[kir] = Кыз досуна кат жазды.
translit[kir] = Qız dosuna qat jazdı.
text[uz] = Qiz do‘stiga xat yozdi.
glossing = girl friend-POSS.3SG-DAT letter write-PST.3SG
text[en] = The girl wrote a letter to her friend.
issue: obl vs. iobj



‘The girl wrote a letter to her friend.’

Figure 4: Example annotation of parallel sentences in four languages.