Unifying the Annotations in Turkic Universal Dependencies Treebanks

Furkan Akkurt[1], Bermet Chontaeva[2], Çağrı Çöltekin[2], Mehmet Oguz Derin, Gulnura Dzhumalieva[3],
Soudabeh Eslami[2], Tunga Güngör[1], Sardana Ivanova[4], Murat Jumashev, Aida Kasieva[3],
Aslı Kuzgun[5], Büşra Marşan[6], Balkız Öztürk[1], Chihiro Taguchi[7],
Susan Üsküdarlı[1], Jonathan Washington[8], and Olcay Taner Yıldız[9]

[1]Boğaziçi University [2]University of Tübingen [3]Kyrgyz-Turkish Manas University
[4]University of Helsinki [5]Starlang Software [6]Stanford University
[7]University of Notre Dame [8]Swarthmore College [9]Özyeğin University

Relevant UniDive working groups: WG1, WG3, WG4

1 Introduction

As the number of treebanks for a single language or a language family in the Universal Dependencies (UD) repository [1] grows, consistent annotations become a concern (Gamba and Zeman, 2023a,b; Zeldes and Schneider, 2023). We report on ongoing efforts to unify Universal Dependencies (UD) treebanks for Turkic languages, currently numbering at 16 in 8 different UD languages. Issues regarding the consistency of UD annotation of Turkic languages have been reported in earlier studies (Tyers et al., 2017; Türk et al., 2019; Çöltekin et al., 2022), with the main consensus being the need for more unified and consistent annotations across treebanks.

The present efforts reported here are coordinated around a one-day workshop alongside the UniDive WG3 meeting in Istanbul in September 2023. Before the meeting, the participants were asked to list difficult-to-annotate linguistic phenomena and inconsistent annotations across existing treebanks. Based on responses, a small example corpus demonstrating these phenomena was prepared [2] and discussed during the workshop and in semi-regular online meetings afterward. In the remainder of this document, we briefly describe a selected subset, and provide a demonstrative example in more detail. We believe the discussion of these linguistic phenomena is likely to increase the consistency of current treebanks, help people preparing new treebanks for Turkic languages (and others facing similar issues), and may result in improvements of the general UD guidelines by demonstrating issues that are not well covered by the current guidelines.

2 Annotation issues in UD Turkic treebanks

Here, we briefly list some recurring issues exhibiting divergence across UD Turkic annotation projects, with the final example discussed in depth.

Tokenization Tokenization, particularly delineating ‘syntactic words’, is a source of inconsistency across Turkic treebanks, especially under differing orthographies for certain clitics and particles (e.g., the question particle). For some historical texts, the lack of uniform word delimiters complicates tokenization as in Old Turkic script, requiring inference via morphosyntactic analysis. Furthermore, some treebanks treat multiple orthographic words as a single token.

Morphological Feature specification There is a proliferation of camel-case tense/mood specifications that could stand to be unified; TAME features in general are applied inconsistently.

Oblique–object distinction Non-accusative objects in Turkic languages are annotated inconsistently, sometimes with the `obj` relation and other times with the `obl` relation.

A strictly morphological perspective favors the oblique analysis of non-accusative objects. This viewpoint is supported through passivization patterns: Non-accusative objects retain their case markings even when they are in the subject position in passive constructions (in the few Turkic languages that even allow promotion of non-accusative objects to subject position), not being marked with nominative, the canonical “subject” case.

[1] See Appendix A for information on current and upcoming Turkic UD treebanks.

[2] Originally in Turkish, but also translated to Azerbaijani, Kyrgyz, Kumyk, Tatar and Old Turkish. Translation and further efforts can be found in our public repository: `github.com/ud-turkic/udtw23`.

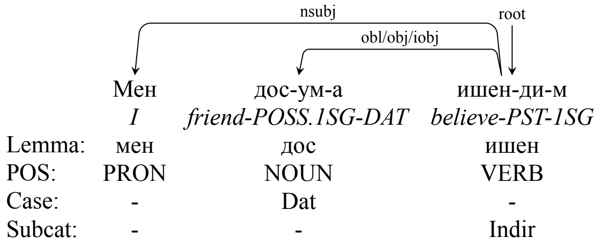


Figure 1: Choice between `obl`, `obj`, and `iobj` relations for a non-accusative object.

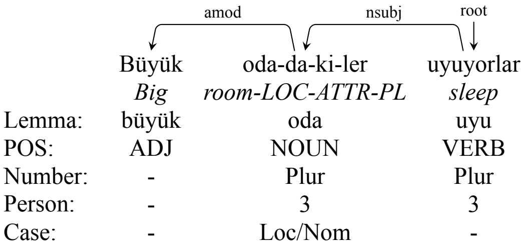


Figure 2: Analyzing *-ki* with no segmentation.

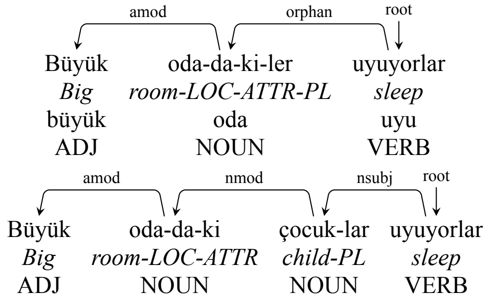


Figure 3: Analysis with orphan (top) and alternative sentence with full noun phrase (bottom).

Conversely, the argument for annotating non-accusative objects with `obj` is rooted in information structure and argument realization. Verbs that typically assign lexical case, such as “to believe” (which assigns dative case in Turkish and Kyrgyz), are transitive. The omission of a dative-marked object in context-free utterances results in a degradation of the utterance. This is in contrast to the omission of true obliques and/or adjuncts, which does not yield a similar effect.

It may alternatively be appropriate to use the `iobj` relation for non-accusative objects like these. The UD v2 documentation suggests that language-specific decisions may be made to annotate sole oblique objects this way.

These three options are presented in Figure 1 with the Kyrgyz sentence *Мен досума ишендим* ‘I believed my friend.’

Question particle Because the question particle functions roughly as an infix in Turkish and is separated by space, various approaches are available for tokenization and part-of-speech annotation.

Code-switching Code-switching, which is a common practice among speakers in many Turkic-speaking multilingual settings, poses challenges for annotation.

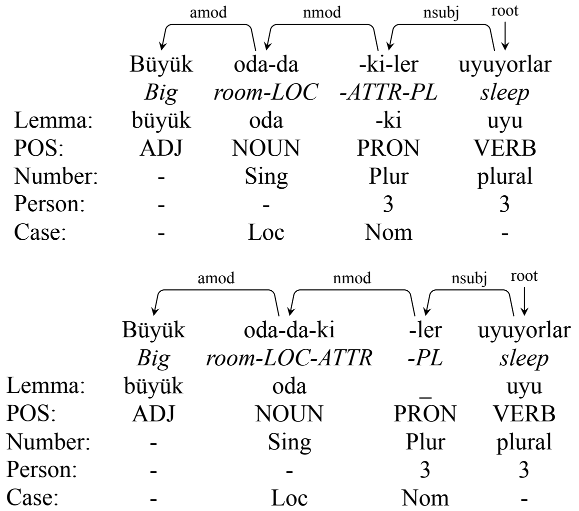
Transcription `Translit` of `MISC` attributes in UD can cover transcription, where coexisting mainstream schemes challenge unified treatment. For example, the Old Turkic script 𐰚𐰃 ‘two’ is transcribed as *äki*, *əki*, *æki*, *eki*, *ėki*, *iki*, *ki*, etc.

Pronominalized locative and genitive nouns Turkic languages allow formation of pronominals from locative and genitive nouns (through the use of *-ki* morpheme in Turkish, *-GI* morpheme in Kyrgyz, etc.). The annotation of the resulting forms is not straightforward. Here we demonstrate four approaches to annotating these forms, and discuss pros and cons of each. We will use the Turkish sentence *Büyük odada**ki**ler uyuyorlar* ‘**The ones** in the big room are sleeping’. The pronominal here may refer to a group of people, e.g., children.

The first option, with **no segmentation** of the word *odadakiler* ‘the ones in the room’ is presented in Figure 2. The advantage of this choice is practical: sub-word segmentation is a non-trivial task, and avoiding it will help making automated analysis more precise, especially in low resource settings. On the other hand, this choice fails to capture that the adjective modifies the noun *oda* ‘room’ (rather than the pronominal), and the lemma `oda` is not the lemma of the subject of the predicate. To complicate matters further, we also fail to capture that the noun is singular and locative, while the resulting pronominal is plural and nominative.[3] In summary, there is a strong indication that the pronominal formed by *-ki* contains multiple syntactic words.

Since the no-segmentation analysis is misleading, another possibility is an analysis with the **`orphan` relation**. This analysis is parallel to that of the sentence *Büyük odadaki çocuklar uyuyorlar* ‘The children in the big room are sleeping’ where a head noun is present, as demonstrated in Figure 3. Although this does not cause misleading/conflicting annotations, the `orphan` analysis is not informative. Furthermore, this analysis

[3] All other `Number` options *oda-lar-da-ki* ‘the one in the rooms’, *oda-lar-da-ki-ler* ‘the ones in the rooms’, and *oda-da-ki* ‘the one in the room’ are also possible.



	Büyük	oda-da	-ki-ler	uyuyorlar
	Big	*room-LOC*	*-ATTR-PL*	*sleep*
Lemma:	büyük	oda	-ki	uyu
POS:	ADJ	NOUN	PRON	VERB
Number:	-	Sing	Plur	plural
Person:	-	-	3	3
Case:	-	Loc	Nom	-

	Büyük	oda-da-ki	-ler	uyuyorlar
	Big	*room-LOC-ATTR*	*-PL*	*sleep*
Lemma:	büyük	oda	_	uyu
POS:	ADJ	NOUN	PRON	VERB
Number:	-	Sing	Plur	plural
Person:	-	-	3	3
Case:	-	Loc	Nom	-

Figure 4: Possible analyses with segmentation, segmenting before *-ki* (top) or after *-ki* (bottom).

does not solve the issue with multiple `Number` or `Case` features that need to be assigned to the noun *odadakiler*.

Segmentation of the pronominalized forms solves the problems with conflicting features and dependencies, as well as the non-informativeness of the `orphan` relation. We consider two different ways (or locations) for segmenting these forms. The first option (Figure 4, top), which is used in some of the current treebanks, considers the *-ki* morpheme as part of the second token, while the second (Figure 4, bottom) segments the word after *-ki*. The first option considers *-ki* as part of the pronominal 'word', which is clearly wrong when considering its attributive use. The problem with the second option is the empty form and lemma when there are no additional affixes after -*ki*. Although it is more principled, this clearly goes against the current UD guidelines.

3 Concluding remarks

We currently do not offer clear recommendations to the issues listed and exemplified above. However, we hope to get a consensus on at least some of the difficult and inconsistent annotations brought up by the Turkic UD community. Some of these issues are relevant to the UD (and UniDive) community at large. As a result, awareness and discussion of these issues may ease new annotation projects for languages with these phenomena, and it is likely to improve the overall quality of the corpora and annotation guidelines.

Acknowledgements

We thank UniDive, the COST Action CA21167, and Istanbul Technical University for supporting the organization of the workshop in September 2023 that made the present work and future collaboration towards a unified annotation of Turkic UD treebanks possible.

References

Ibrahim Benli. 2023. UD Kyrgyz KTMU treebank. GitHub repository.

Neslihan Cesur, Aslı Kuzgun, Olcay Taner Yıldız, Büşra Marşan, Neslihan Kara, Bilge Nas Arıcan, Merve Özçelik, and Deniz Baran Aslan. 2021a. UD Turkish Penn treebank. GitHub repository.

Neslihan Cesur, Aslı Kuzgun, Olcay Taner Yıldız, Büşra Marşan, Oğuzhan Kuyrukçu, Bilge Nas Arıcan, Ezgi Sanıyar, Neslihan Kara, and Merve Özçelik. 2021b. UD Turkish FrameNet treebank. GitHub repository.

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.

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 Doğruöz, and Özlem Çetinoğlu. 2022. Resources for Turkish natural language processing: A critical survey. *Language Resources and Evaluation*.

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.

Federica Gamba and Daniel Zeman. 2023a. Latin morphology through the centuries: Ensuring consistency for better language processing. In *Proceedings of the Ancient Language Processing Workshop*, pages 59–67, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Federica Gamba and Daniel Zeman. 2023b. Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 7–16, Washington, D.C. Association for Computational Linguistics.

Aida Kasieva, Gulnura Dzhumalieva, Anna Thompson, Murat Jumashev, Bermet Chontaeva, and Jonathan Washington. 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)*.

Aslı Kuzgun, Neslihan Cesur, Olcay Taner Yıldız, Oğuzhan Kuyrukçu, Büşra Marşan, Bilge Nas Arıcan, Neslihan Kara, Deniz Baran Aslan, Ezgi Sanıyar, and Cengiz Asmazoğlu. 2021a. UD Turkish tourism treebank. GitHub repository.

Aslı Kuzgun, Neslihan Cesur, Olcay Taner Yıldız, Oğuzhan Kuyrukçu, Arife Betül Yenice, Bilge Nas Arıcan, and Ezgi Sanıyar. 2021b. UD Turkish Kenet treebank. GitHub repository.

Mehmet Köse and Olcay Taner Yıldız. 2021. UD Turkish Atis treebank. GitHub repository.

Aibek Makazhanov, Aitolkyn Sultangazina, Olzhas Makhambetov, and Zhandos Yessenbayev. 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)*, pages 338–350.

Büşra Marşan, Salih Furkan Akkurt, Muhammet Şen, Merve Gürbüz, Onur Güngör, Şaziye Betül Özateş, Suzan Üsküdarlı, Arzucan Özgür, Tunga Güngör, and Balkız Öztürk. 2022. Enhancements to the boun treebank reflecting the agglutinative nature of turkish. In *The Proceedings of the ALTNLP2022 The International Conference and workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing*, pages 71–80.

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, Marseille, France. European Language Resources Association.

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. 2021. UD Tatar NMCTT treebank. GitHub repository.

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, Paris, France. Association for Computational Linguistics.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Gözde Berk, Seyyit Talha Bedir, Abdullatif Köksal, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2022. Resources for turkish dependency parsing: Introducing the boun treebank and the boat annotation tool. *Language Resources and Evaluation*, pages 1–49.

Francis Tyers and Jonathan Washington. 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*, pages 276–289.

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.

Amir Zeldes and Nathan Schneider. 2023. Are UD treebanks getting more consistent? a report card for English UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 58–64, Washington, D.C. Association for Computational Linguistics.

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, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 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 UD Turkic Treebanks

There are currently UD treebanks for Kazakh, Kyrgyz, Tatar, Turkish, Uyghur, Yakut, and Old Turkish, and a treebank annotating sentences with Turkish-German code switching. All languages except Turkish are represented with a single treebank, while Turkish has 9 treebanks. Table 1 lists the treebanks currently released in the UD repositories as of UD version 2.13.

	sent	tok	multi	types	ltypes	pos	rel	feat
Kazakh/KTB (Tyers and Washington, 2015; Makazhanov et al., 2015)	1078	10536	41	4642	2433	17	36	9
Kyrgyz/KTMU (Benli, 2023)	781	7451	0	3474	2305	13	26	8
Old Turkish/Tonqq (Derin and Harada, 2021)	20	158	0	75	2	13	19	0
Tatar/NMCTT (Taguchi, 2021)	148	2280	0	1264	843	14	28	7
Turkish/Atis (Köse and Yıldız, 2021)	5432	45907	0	2133	995	13	36	7
Turkish/BOUN (Türk et al., 2022; Marşan et al., 2022)	9761	125212	3374	37052	12649	16	46	7
Turkish/FrameNet (Cesur et al., 2021b)	2698	19223	0	8403	3905	15	30	7
Turkish/GB (Çöltekin, 2015)	2880	17177	371	5517	2074	16	42	7
Turkish/IMST (Sulubacak et al., 2016)	5635	58096	1639	18541	5960	14	40	10
Turkish/Kenet (Kuzgun et al., 2021b)	18687	178658	0	49156	15343	15	34	7
Turkish/Penn (Cesur et al., 2021a)	16396	183555	0	37765	14977	15	36	9
Turkish/PUD (Zeman et al., 2017)	1000	16881	346	7646	4598	16	38	4
Turkish/Tourism (Kuzgun et al., 2021a)	19830	91152	0	4961	2170	15	33	13
Turkish-German/SAGT (Çetinoğlu and Çöltekin, 2022)	2184	37227	290	7094	3836	17	45	12
Uyghur/UDT (Eli et al., 2016)	3456	40236	0	12067	2908	16	45	15
Yakut/YKTDT (Merzhevich and Ferraz Gerardi, 2022)	299	1460	1	688	405	14	26	6

Table 1: Basic statistics on current UD treebanks (as of UD version 2.13). *sent*: number of sentences, *tok*: number of tokens, *multi*: number of multi-word tokens, *types*: number of word types, *ltypes*: number of lemma types, *pos*: number of POS tags used, *rel*: number of dependency relations used (including language/treebank specific relations), *feat*: number of morphological features used.

Besides existing treebanks, the UD web page also reports Uzbek, Ottoman Turkish and yet another Turkish treebank in preparation. We are also aware of new treebanks in preparation for Kyrgyz (Kasieva et al., 2023), Azerbaijani and Kumyk.