Tokenisation of Turkic Copula Constructions in Universal Dependencies

Çağrı Çöltekin[1], **Furkan Akkurt**[2], **Bermet Chontaeva**[1], **Soudabeh Eslami**[1], **Sardana Ivanova**[3], **Gulnur Dzhumalieva**[4], **Aida Kasieva**[4], **Nikolett Mus**[5], **Jonathan Washington**[6]

[1]University of Tübingen, [2]Boğaziçi University, [3]Independent Researcher,

[4]Kyrgyz-Turkish Manas University, [5]Hungarian Research Centre for Linguistics, [6]Swarthmore College

cagri.coeltekin@uni-tuebingen.de, furkan.akkurt@bogazici.edu.tr, bermet.chontaeva@student.uni-tuebingen.de, soudabeh.eslami@student.uni-tuebingen.de, sardana.n.ivanova@gmail.com, gulnur.jumalieva@manas.edu.kg, aida.kasieva@manas.edu.kg, mus.nikolett@nytud.hun-ren.hu, jonathan.washington@swarthmore.edu

Abstract

Identifying units, 'syntactic words', for morphosyntactic analysis is important yet challenging for morphologically rich languages. In this paper we propose a set of guiding principles to determine units of morphosyntactic analysis, and apply them to the case of copular constructions in Turkic languages, in the context of Universal Dependencies (UD) framework. We also provide a survey of the practice in the Turkic UD treebanks published to date, and discuss the advantages and disadvantages of the proposed tokenisation for a selection of Turkic languages.

1 Introduction

The linguistic unit *word* is central to linguistic analysis and description of a language. Yet, it has been very difficult to come up with a clear definition (Haspelmath, 2023).[1] Defining the word becomes even more difficult for morphologically rich languages and language families, where the boundary between a morpheme and word is often unclear since information encoded by multiple words in other languages can be encoded within a single word. The boundary can further be blurred by language change, where a 'full word' may become a clitic, and further an affix over time.

How wordhood is defined affects both linguistic and computational analyses, determining both simplicity and elegance of linguistic analysis, and success of natural language processing (NLP) applications. A consistent definition across related languages may further aid cross-linguistic studies and cross-lingual transfer in NLP applications, particularly enabling effective use of linguistic resources in related (high-resource) languages in NLP applications for low-resource languages. In this paper we propose an analysis for an interesting case of unclear word boundaries, copular constructions in Turkic languages, from the perspective of the Universal Dependencies (UD, De Marneffe et al., 2021).

The UD project defines a unified morphological and syntactic annotation framework for dependency analysis of natural languages. As well as its clear contributions to quantitative and comparative linguistic research, the UD treebanks are a valuable source for training models for extremely low-resource languages (Gessler and Zeldes, 2022), and they also provide a valuable resource for building linguistically-informed benchmarks (Linzen et al., 2016; Warstadt et al., 2020; Başar et al., 2025). Furthermore, even though current pretrained models replaced the traditional NLP pipelines with an end-to-end model, (subword) tokenisation is still the initial step that all current models have to perform. As a result, linguistically-informed approaches to tokenisation are also likely to improve performance of (large) language models as well as reducing the language bias in multilingual models (Petrov et al., 2023; Toraman et al., 2023).

Universal Dependencies allow the use of so-called multi-word tokens, where an 'ortho-

[1]What 'looks like' a word also differs based on the linguistic analysis of interest. One often encounters specialized definitions like *phonological word* or *syntactic word*. In this work, we are primarily concerned with the syntactic word.

graphic word' can be tokenised into multiple 'syntactic words', the unit of morphosyntactic analysis in UD. Although there are no clear principles or guidelines that define a syntactic word in UD, many languages make use of this mechanism to use units smaller than the orthographic word as the basic unit of morphosyntactic analysis. The need for such units was also recognized in earlier work in Turkish CL, where it was customary to tokenise orthographic words into units that are called *inflectional groups* (IG, Oflazer, 2003; Çöltekin, 2016). The first Turkic dependency treebank (that we are aware of), Turkish METU-SABANCI treebank, (Say et al., 2002; Oflazer et al., 2003) also uses IGs as the unit of syntactic analysis. The IG was loosely defined as a unit within a word with its own inflections. This typically includes all productive derivational affixes, as well as affixes that change the grammatical function of the word in some way, resulting in a rather large number of inflectional groups as exemplified in Figure 1.

Current Turkic UD treebanks follow a more conservative tokenisation strategy. However, the approach used in each treebank, even within the same language, differs substantially (see Section 3 for details). The choice of the right level of tokenisation is also a part of an ongoing discussion in the broader UD community (Guillaume et al., 2024; Evang and Zeman, 2024). In this paper, we focus on the tokenisation of the copular constructions in Turkic Universal Dependencies treebanks. The main principle behind our proposal is the *lexical integrity* principle, which states "Rules of syntax can refer/apply to entire words or the properties of entire words, but not to the internal parts of words or their properties." (Haspelmath and Sims, 2010, p.203). As well as being an initiative towards more uniform tokenisation and annotation of copular constructions across Turkic UD treebanks, the present paper also aims to inform efforts to establish tokenisation guidelines in the broader context.

In the remainder of this paper, we provide a brief survey of the copular constructions in a sample of Turkic languages and an overview of the different ways of annotating these constructions in existing Turkic UD treebanks. We define the problem, and propose our solution in Section 4, and provide a summary and future directions in Section 5.

Language	1SG	2SG	3SG	1PL
Azerbaijani	-(y)Am	-(s)An	(-DIr)	-(y)IK
Kazakh	-MIn	-sIŋ	–	-MIz
Kyrgyz	-mIn	-sIŋ	–	-BIz
Qaraqalpaq	-MAn	-sAŋ	–	-MIz
Sakha	-BIn	-GIn	–	-BIt
Tatar	-mIn	-sIŋ	–	-BIz
Turkish	-(y)Im	-sIn	–	-(y)Iz
Tuvan	men	sen	–	bis
Uzbek	-man	-san	(-DIr)	-miz

Table 1: A selection of copula agreement suffix forms in some Turkic languages. - = morpheme boundary within a word; capital letters = realised in multiple ways; () = realised only in some phonological environments or optional.

2 Copular Constructions in Turkic Languages

Turkic languages show a range of copular constructions, from full words and free morphemes to fully bound morphological units. Furthermore, no copular affix or verb is expressed at all in some constructions. For example, Old Turkic present copula sentences (1a) optionally use a full copular verb *är*, with agreement as a stand-alone echoing the pronoun either way, while the equivalent Turkish example (1b) expresses the same predication without an overt verb and agreement is expressed as morphological suffixes attached to the non-verbal predicate. In this case, *-yim*, which diachronically derives from a first-person pronoun, is attached to a locative noun.

(1) a. *Bän äbdä (är-ür) män.*
I house-LOC (be-NPST) 1SG
'I am at home.'

b. *Ben evdeyim.*
I house-LOC-1SG
'I am at home.'

Whether a full word, affix, or both can be used in copular constructions varies across Turkic languages. In this paper, we are concerned with non-finite agreement marking, which is often bound (Table 1), as well as bound forms of the copula verb ER (Table 2).[2]

[2]ER refers to Turkic copula verbs that are understood to descend from proto-Turkic **är* 'be'.

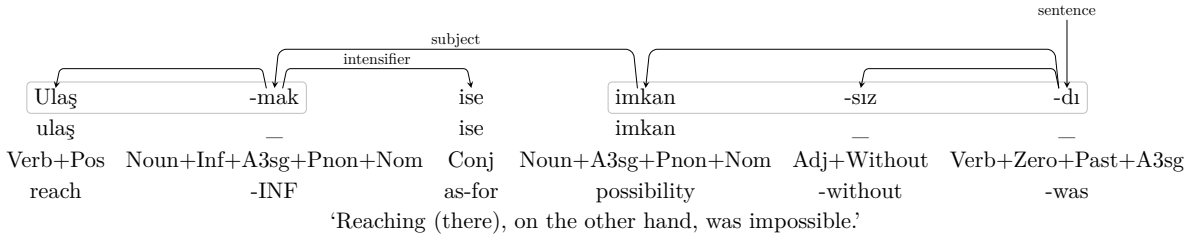

Figure 1: A dependency analysis from the METU-SABANCI treebank. Forms for the affixes are added for clarity. The treebank does not specify the forms and lemmas of non-root IGs. The second row is the lemmas in the dependency analysis, and the third row provides the native morphological analysis as annotated in the treebank, and the last row is the gloss. The treebank does not specify labels for within-word dependencies.

Language	-DI		-mIş		-sA		VN
	free	bound	free	bound	free	bound	
Azerbaijani	idi	-(y)DI	imiş	-(y)mIş	isə	-(y)sA	–
Kazakh	еді	–	–	–	–	–	екен(діг)і
Kyrgyz	эле	–	–	–	–	–	экен(диг)и
Sakha	этэ	–	эбит	–	–	–	–
Tatar	иде	–	–	–	–	–	икән(лег)е
Turkish	idi	-(y)DI	imiş	-(y)mIş	ise	-(y)sA	–
Uzbek	edi	–	–	–	isa	–	ekan(lig)i

Table 2: Attested free and bound forms of the defective copula verb ER across a selection of modern Turkic languages. Some forms have alternative uses, including as topicalisers, conjunctions, or sentence- or clause-level markers of epistemic modality. VN = verbal noun.

3 An Overview of Current practice in Turkic UD Treebanks

A survey of current Turkic treebanks shows a variety of strategies used to annotate and tokenise copula constructions. The two main strategies for bound forms of the copula verb and bound agreement suffixes were either to tokenise them as subtokens or to put agreement or verbal features on non-verbal predicates. A partial summary of our survey is included in Appendix A.

4 Problems and the Proposed Solution

As noted in Section 1, our main guiding principle for tokenisation of the copular constructions is the lexical integrity hypothesis. Within the context of morphosyntactic dependency analysis, we explicate it as, (i) no syntactic relation should refer to parts of a syntactic word; and (ii) the syntactic unit should not require repeated or conflicting morphological features. Besides these guiding principles, we also consider the following properties desirable for a definition of syntactic word for a morphosyntactic analysis/annotation framework: (iii) compliant with the main principles of the framework (UD); (iv) a close match with orthography of the language; (v) similar analysis of similar constructions within a language as well as across (related) languages; but ideally (vi) no loss of information, or bleaching of linguistic distinctions or differences between languages. In this section, we first demonstrate the problems with a no-segmentation approach, where the orthographic word boundaries are followed. Then, we outline the proposed segmentation strategy.

Table 2 lists the means of expressing copular constructions using the copula verb in a selection of Turkic languages, where the verb form may be free or bound. Example (2), in Turkish, demonstrates both possibilities, where (2a) makes use of the free copula verb *i-*, while in (2b), the copula verb is expressed as a bound morpheme *-y-*.

(2) a. *Ben dağlarda idim.*
I mountain-PL-LOC COP-PST-1SG
'I was in the mountains.'

b. *Ben dağlardaydım.*
I mountain-PL-LOC-COP-PST-1SG
'I was in the mountains.'

In cases where a full verb is used (2a), the analysis is straightforward without further segmentation. For the cases where copula is expressed as an affix, the no-segmentation approach violates the lexical integrity principle. Notice that in (2b), the same word includes feature PL, and the noun could be considered third person (3), which would be translated to Number=Plur|Person=3 in UD, and 1SG translates to Number=Sing|Person=1 in UD, leading

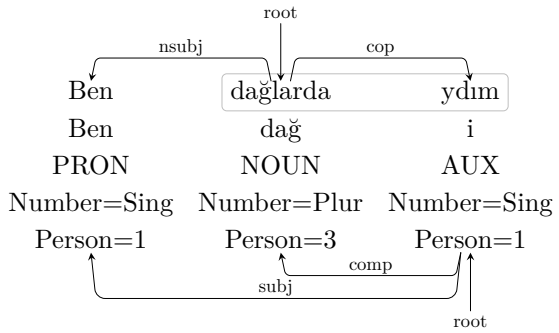

Figure 2: The proposed dependency analysis of example (2b). Both UD (above) and SUD (below) dependencies are shown. Second row: lemmas, third row POS tags, fourth and fifth rows morphological features (only ones relevant to the discussion are shown).

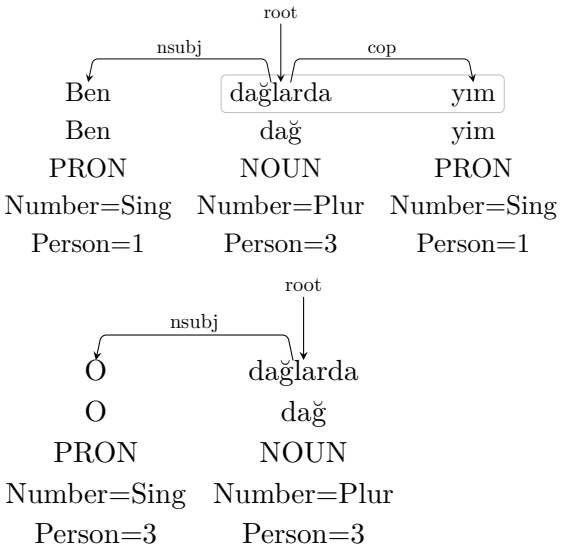

Figure 3: UD analysis of present tense copula, with first person (top) and third person (bottom) forms

to a violation of principle (ii). Furthermore, the word *dağlardaydım* functions as both a noun and a verb (predicate) in (2b), also clearly violating principle (ii). For UD, this does not cause a conflict with principle (i), since UD allows nouns to act as predicates, and even when expressed by a verb, copular verbs are marked as dependents of the complement. However, for related formalisms like Surface Syntactic Universal Dependencies (SUD, Gerdes et al., 2018) where a copula has to be marked as a head of the construction, a no-segmentation approach would also violate principle (i) above.

The solution we propose is tokenising the affixal copular constructions, which leads to an analysis as shown in Figure 2. This analysis does not violate principles (i) and (ii) – all dependencies involve syntactic words, there are no repeated or conflicting morphological features. Furthermore, the solution does not offer excessive segmentation (complies with iii); provides similar analyses for similar constructions within the language family (complies with v); marks the difference between full verb and affix/clitic version through UD multi-word tokens (complies with vi); trading off (iv) by requiring a non-trivial segmentation of the orthographic words.

Turkic present-tense copular constructions with bound agreement marking present a challenge for this proposal. Unlike other copular constructions, in these constructions there is no clear copular verb present. For example, despite surface resemblance, the segment *-y-* in (1b) is not an affix (unlike the same segment in (2b)). The distinction is further evidenced by the fact that (1b) does not have a full-verb copula version, **Ben evde im* is not grammatical. What follows the nominal are only person markers. Nevertheless, the problems discussed for bound copula verb forms are still relevant to the present copula. For example, the present tense version of (2b), *Ben dağlardayım*, still requires conflicting morphological features on a single word. As a result, we propose tokenising the person markers attached to non-verbal predicates in present tense copular constructions.

One downside of this analysis is the inconsistency between similar structures as shown in Figure 3. Two sentences that are identical—except for person markers—receive different dependency analyses, one with and one without copula dependency. A possible solution is to introduce syntactic words without surface forms. However, since null elements are not allowed in the UD basic dependencies (like many other computational syntactic formalisms), we believe the current proposal still is a good compromise. The treebanks that make use of UD enhanced dependencies can still mark the similarity between these two sentences in the enhanced dependency annotation.

This leaves the question of how to annotate the lemma of bound forms of ER and bound agreement marking. Treebanks that tokenise bound ER are fairly consistent in using the same lemma as free ER (such as 'i' in Turk-

ish). Some treebanks that tokenise bound agreement marking use the same lemma for these subtokens. Other possibilities are to use the lemma of the corresponding pronoun (such as 'ben' for *-(y)Im* in Turkish) or to use a single consistent form of the agreement morpheme (such as 'yım' for *-(y)Im*).

5 Conclusions and Outlook

We have proposed guidelines for UD annotation of bound copular constructions in Turkic languages. Specifically, we propose separate tokenisation of bound copula agreement affixes and bound forms of the copula verb ER. This solves issues of conflicting morphological features despite not aligning with orthographic notions of wordhood. In addition, and perhaps at least as importantly, we have laid out principles that we feel are applicable to any decision about how to tokenise and annotate a construction in UD.

Future work is needed to address a number of adjacent issues, including copular negation, copular forms used with non-finite verb forms, verbal noun forms of copular verbs, the *-DIr* morpheme that occurs in copular constructions in some Turkic languages, other copula strategies such as the full verb BOL (addressed in Kasieva et al., 2025), and how tokenisation of the question particle interacts with copula tokenisation.

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.

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.

Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. 2025. TurBLiMP: A Turkish benchmark of linguistic minimal pairs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16495–16510, Suzhou, China. Association for Computational Linguistics.

İbrahim Benli. 2021. UD Kyrgyz-KTMU treebank. UD version 2.17, November 15, 2025.

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. 2016. (When) do we need inflectional groups? In *Proceedings of The First International Conference on Turkic Computational Linguistics*.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Kilian Evang and Daniel Zeman. 2024. Word segmentation in Universal Dependencies. In *Second UniDive General Meeting*.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium. Association for Computational Linguistics.

Luke Gessler and Amir Zeldes. 2022. MicroBERT: Effective training of low-resource monolingual BERTs through parameter reduction and multitask learning. In *Proceedings of the 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 86–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Bruno Guillaume, Kim Gerdes, Kirian Guiller, Sylvain Kahane, and Yixuan Li. 2024. Joint annotation of morphology and syntax in dependency treebanks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9568–9577, Torino, Italia. ELRA and ICCL.

Martin Haspelmath. 2023. Defining the word. *Word*, 69(3):283–297.

Martin Haspelmath and Andrea D. Sims. 2010. *Understanding Morphology*, 2e edition. Understanding language series. Hodder Education.

Aida Kasieva, Nikolett Mus, Arofat Akhundjanova, Furkan Akkurt, Bermet Chontaeva, Gulnura Dzhumalieva, Soudabeh Eslami, Murat Jumashev, and Jonathan Washington. 2025. The annotation of Turkic copula verbs in UD. In *2025 10th International Conference on Computer Science and Engineering (UBMK)*, pages 1772–1777.

Aslı Kuzgun, Oğuz Kerem Yıldız, Neslihan Cesur, Büşra Marşan, Arife Betül Yenice, Ezgi Sanıyar, Oguzhan Kuyrukçu, Bilge Nas Arıcan, and Olcay Taner Yıldız. 2021a. From constituency to UD-style dependency: Building the first conversion tool of Turkish. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 761–769, Held Online. INCOMA Ltd.

Aslı Kuzgun, Neslihan Cesur, Olcay Taner Yıldız, Oğuzhan Kuyrukçu, Arife Betül Yenice, Bilge Nas Arıcan, and Ezgi Sanıyar. 2021b. UD Turkish-Kenet treebank. UD version 2.17, November 15, 2025.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Aibek Makazhanov, Aitolkyn Sultangazina, Olzhas Makhambetov, and Zhandos Yessenbayev. 2015. Syntactic annotation of Kazakh: Following the universal dependencies guidelines. a report. In *3rd International Conference on Turkic Languages Processing, (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. Enhancements to the BOUN Treebank Reflecting the Agglutinative Nature of Turkish.

Sanatbek Matlatipov and Elmurod Kuriyozov. 2025. UD Uzbek UzUDT treebank. GitHub repository.

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.

Kemal Oflazer. 2003. Dependency parsing with an extended finite-state approach. *Computational Linguistics*, 29(4):515–544.

Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür. 2003. Building a Turkish treebank. In *Treebanks: Building and using parsed corpora*, pages 261–277. Springer.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *Advances in neural information processing systems*, 36:36963–36990.

Bilge Say, Deniz Zeyrek, Kemal Oflazer, and Umut Özge. 2002. Development of a corpus and a treebank for present-day written Turkish. In *Proceedings of the eleventh international conference of Turkish linguistics*, pages 183–192. Eastern Mediterranean University.

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.

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahinuç, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for turkish. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–21.

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.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

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 A Survey of Tokenisation of Copula in Turkic UD Treebanks

Treebank	Source	Agreement suffixes	Bound ER
Azerbaijani-TueCL	Akhundjanova et al. (2025)	N/A	subtoken, lemma=i
Kazakh-KTB	Tyers and Washington (2015); Makazhanov et al. (2015)	subtoken, lemma=e	N/A
Kyrgyz-KTMU	Benli (2021)	not tokenised	N/A
Tatar-NMCTT	Taguchi et al. (2022)	not tokenised	N/A
Turkish-BOUN	Marşan et al.	subtoken, lemma=i	subtoken, lemma=y
Turkish-GB	Çöltekin (2015)	subtoken, lemma=i	subtoken, lemma=i
Turkish-Kenet	Kuzgun et al. (2021b)	not tokenised	not tokenised
Turkish-Penn	Kuzgun et al. (2021a)	not tokenised	not tokenised
Turkish-PUD	Zeman et al. (2017)	subtoken, lemma=i,_	subtoken, lemma=i
Turkish-TueCL	Akhundjanova et al. (2025)	N/A	subtoken, lemma=i
Uzbek-UDT	Matlatipov and Kuriyozov (2025)	not tokenised	N/A
Uzbek-UT	Akhundjanova and Talamo (2025)	not tokenised	N/A
Yakut-YKTDT	Merzhevich and Ferraz Gerardi (2022)	not tokenised	N/A

Table 3: A survey of copula tokenisation for a selection of Turkic UD treebanks (v.2.17). subtoken = an extra subtoken with its own features; not tokenised = no subtokenisation, copula features on last token of predicate; N/A = not attested in corpus (or language). Most treebanks include inconsistencies. In case we observe both tokenisation and no tokenisation in the same treebank, we assume tokenisation is the intended behaviour.