# Universal Dependencies Treebank for Uzbek

**Arofat Akhundjanova** and **Luigi Talamo**
Saarland University / Saarbrücken, Germany
arak00001@stud.uni-saarland.de, luigi.talamo@uni-saarland.de

## Abstract

We present the first Universal Dependencies treebank for Uzbek, a low-resource language from the Turkic family. The treebank contains 500 sentences (5850 tokens) sourced from the news and fiction genres and it is annotated for lemmas, part-of-speech (POS) tags, morphological features, and dependency relations. We describe our methodology for building the treebank, which consists of a mix of manual and automatic annotation and discuss some constructions of the Uzbek language that pose challenges to the UD framework.

## 1 Introduction

Although Uzbek ranks as the second Turkic language in terms of speakers after Turkish (Boeschoten, 2021a), computational resources for this language are scarce. We aim to partially fill this gap by introducing the first fully annotated Universal Dependencies (UD) treebank for Uzbek - Uzbek-UT (Uzbek Universal Treebank)[1].

The UD framework facilitates consistent morpho-syntactic annotation across different languages (de Marneffe et al., 2021) and represents an open community initiative aimed at creating annotated corpora for numerous languages. As of v.2.15, UD includes 296 treebanks covering 168 languages[2]. Nowadays treebanks are essential for the development of Natural Language Processing (NLP) tools and are also increasingly used in linguistic research.

The present paper is organized as follows. In Section 2, we provide a brief sketch of the Uzbek language and in Section 3, we review the existing computational resources for Uzbek. Section 4 forms the core of the paper, describing the steps involved in the treebank development, including automatic annotation and manual correction. In Section 5, we analyze some constructions that pose challenges to the UD framework. Section 6 summarizes our work and proposes directions for future research.

## 2 The Uzbek Language

Uzbek is a member of the Karluk branch of the Turkic language family and has the status of official language in Uzbekistan. With over 40 million speakers, it is primarily used in Uzbekistan and surrounding Central Asian countries, and considered as the second-most widely spoken Turkic language after Turkish (Boeschoten, 2021a).

The official script of the language is Latin, but the old Cyrillic script is still in use (Boeschoten, 2021b, 390). The treebank described in this work only contains Uzbek sentences written in the Latin script.

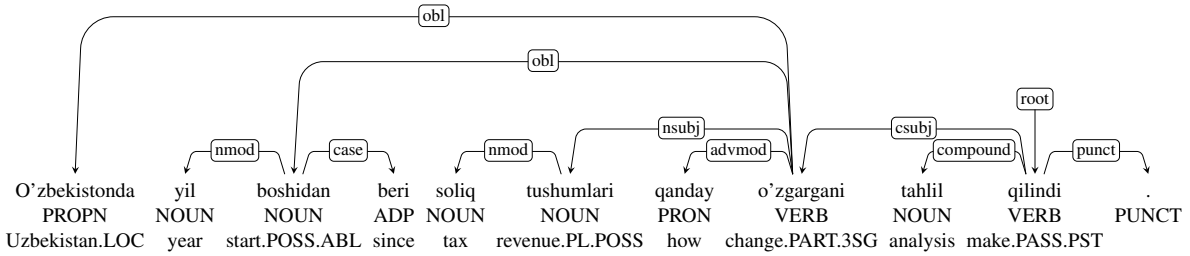
Uzbek grammar shares similarities with other Turkic languages, but computational resources developed for cognate languages cannot be directly applied. From a typological perspective, Uzbek is a null-subject, highly agglutinative language and lacks gender distinctions and articles. Like other Turkic languages, Uzbek has a basic SOV word order, which is quite flexible and can be easily altered for information structure by fronting the topic (Boeschoten, 2021b, 401-407). Its morphology is highly regular and the standard orthography does not indicate vowel harmony or consonant assimilation. Modifiers precede the head noun and are generally follow the pronoun-quantifier-adjective order. Number agreement in the nominal phrase is not obligatory, and nouns modified by quantifiers are often unmarked for plural. (Boeschoten, 2021b, 392-393)

[1]The treebank is available online at https://github.com/UniversalDependencies/UD_Uzbek-UT.

[2]https://universaldependencies.org/



*Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025)*, pages 1–6




'How tax revenues have changed in Uzbekistan since the beginning of the year was analyzed.'

Figure 1: UD annotation of an Uzbek sentence

## 3 Related Work

Early computational resources for Uzbek included a morphological parser written in Prolog (Matlatipov and Vetulani, 2009), which however lacked support for complex words. Sharipov et al. (2022) introduced an expanded tagset through deeper morphological and syntactic analysis. This was followed by the creation of UzbekTagger, a rule-based POS tagger (Sharipov et al., 2023), which was based on 12 POS tags and tested on the manually annotated data.

The development of stemmers and lemmatizers (Sharipov and Yuldashov, 2022; Sharipov and Sobirov, 2022) has been another important contribution. UzMorphAnalyzer, introduced by Salaev (2023), represents a more comprehensive tool, integrating a stemmer, lemmatizer, and POS tagger. Additionally, a robust finite-state transducer (FST)-based morphological analyzer, included in the Apertium monolingual package, supports Uzbek text processing[3].

Significant efforts have also been directed toward dataset creation, including WordNet-type synsets (Agostini et al., 2021; Madatov et al., 2022), sentiment analysis datasets (Kuriyozov et al., 2019; Matlatipov et al., 2022), semantic evaluation dataset (Salaev et al., 2022) and text classification datasets (Rabbimov and Kobilov, 2020; Kuriyozov et al., 2023). However, there remains a lack of a fully annotated gold-standard dataset for training automatic taggers and parsers.

In recent years, neural transformer-based language models like UzBERT (Mansurov and Mansurov, 2021) and BERTbek (Kuriyozov et al., 2024) have emerged. These models were pretrained and evaluated against multilingual BERT (Devlin et al., 2019), showing promising results in masked language modeling and other downstream tasks.

## 4 Treebank Development

### 4.1 Overview and Data Selection

The treebank building consists of the following steps: (i) word segmentation and lemmatization, (ii) morphological and Universal Parts-of-Speech (UPOS) tagging and (iii) dependency parsing. We cover all the annotation fields in the CoNLL-U format[4], except for the language-specific part-of-speech tagset (XPOS) and the enhanced dependency graph (DEPS). Figure 1 shows an Uzbek sentence to exemplify different UD annotation fields.

Our methodology combines automated processing with manual annotation and revision. Whenever possible, processing tasks were performed automatically using existing tools, and then revised manually by a native Uzbek-speaking author with a background in Uzbek linguistics. The entire treebank underwent manual verification and correction to resolve ambiguities, eliminate errors and ensure consistency. Ambiguous cases were solved through extensive discussions with other linguists and UD experts.

The treebank contains 500 sentences (5,850 tokens), 250 of which are collected from news articles and 250 from fiction books. The news sentences are taken from the UzCrawl dataset (Mamasaidov and Shopulatov, 2023), which collected data from major news sites[5] covering diverse topics and representing modern Uzbek language usage. The fiction sentences are selected from the publicly available 20th- and 21st-century Uzbek literary works found online. To ensure data qual-

[3] https://github.com/apertium/apertium-uzb

[4] https://universaldependencies.org/format.html

[5] https://kun.uz/ and https://daryo.uz/

| | Sentences | Tokens | Unique words | POS tags | Features | Dependencies |
|---|---|---|---|---|---|---|
| **No.** | 500 | 5850 | 3523 | 17 | 42 | 32 |

Table 1: Basic statistics for the UT treebank.

| Model run No. | No. of sentences | | | Tokenizer | Lemmatizer | UPOS Tagger | Parser |
|---|---|---|---|---|---|---|---|
| | train | test | dev | | | | |
| **1st run** | 100 | - | 50 | 99.86 | 86.78 | 69.39 | 46.26 |
| **2nd run** | 240 | 30 | 30 | 96.72 | 86.88 | 68.22 | 48.98 |
| **3nd run** | 400 | 50 | 50 | 98.30 | 92.11 | 73.08 | 52.43 |

Table 2: Model evaluation with F1 score for the three runs.

ity, all sentences were manually selected. The inclusion of both news and fiction ensures coverage of different domains, levels of formality, and stylistic variations. The two genres are distinguishable by sentence IDs: the first half of the treebank corresponds to news, while the second half belongs to fiction. Table 1 provides basic statistics for the treebank.

### 4.2 Word Segmentation and Lemmatization

The segmentation of sentences into words was performed automatically with the NLTK tokenizer[6] (Loper and Bird, 2002). The tokenized data amounts to 5,850 tokens. Currently, UD does not permit words containing spaces. Although multiword expressions (MWEs) are conceptually treated as single words, they are annotated using specific dependency relations rather than being merged into a single token. For example, the proper noun *Tog'li Qorabog'* 'Nagorno-Karabakh' is segmented into two tokens and annotated with `flat` relation. Punctuation marks that are attached to a word are tokenized as separate words; exceptions are full stop marking an abbreviation, which are tokenized together, e.g. *mln.* 'million', *A. Navoiy* 'A. Navoi'.

Lemmatization was performed automatically with the UzMorphAnalyzer tool (Salaev, 2023). However, since UzMorphAnalyzer does not disambiguate between identical tokens with different lemmas, manual disambiguation was required.

### 4.3 UPOS and Morphological Tagging

UPOS tagging is notably a tedious and time-consuming task. In order to speed up the annotation process, we tagged the tokens with the UzMorpAnalyser. Before starting the tagging process, we first mapped traditional Uzbek word classes (Rahmatullayev, 2006) to 17 UPOS tags, adhering to the UD guidelines[7].UPOS-tagged tokens were then manually checked and corrected, as the tagger did not reach a satisfactory level of accuracy.

For morphological features, which are referred to as 'Universal features'[8] in the UD framework, we first selected 42 Universal features and annotated 150 sentences manually. We then used these sentences as training data to build a parser for automatically tagging the remaining sentences. For this task, we used Stanford Stanza [9] (Qi et al., 2020), a Python-based NLP library with neural network components. This significantly reduced manual work, as some Universal features were predicted with near-perfect accuracy. As the final step of this task, we manually revised and corrected the annotations for 350 sentences.

### 4.4 Dependency Parsing

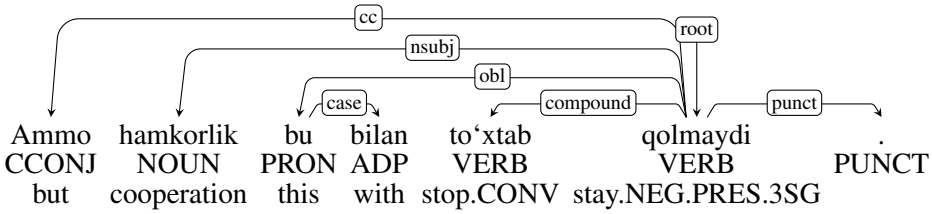
To train a dependency parser, Stanford Stanza requires a pipeline with three interconnected processors: a tokenizer, lemmatizer and POS tagger. Therefore, we left dependency parsing as the last step in building the treebank. We first selected 32 UD syntactic relations and manually annotated 150 sentences with the help of Grew tools (Guillaume, 2021). Together with Uzbek word vectors from the fastText collection (Grave et al., 2018), we used these sentences to train an initial Stanza dependency parsing model (1st run). This model was then used to parse an additional 200 sentences, which were manually corrected for dependency relations and used to train a second model

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

[7] `https://universaldependencies.org/u/pos/index.html`

[8] `https://universaldependencies.org/u/feat/index.html`

[9] `https://stanfordnlp.github.io/stanza/`




‘But the cooperation does not end with this.’

Figure 2: Annotation for the postverbial construction *to’xtab qol*.

(2nd run) Finally, we re-iterated the training and correction process with the remaining sentences to train a final model (3rd run). Table 2 shows the performance improvements over the three runs.

## 5 Challenging Constructions

In this section, we address some of the challenges we have encountered in building the UT treebank for different annotation fields: UPOS, Universal Features and syntactic relations.

As for UPOS tagging, the Particle + Verb pattern used in verbal multi-word expressions (MWEs) is particularly challenging, as the Particle does not have a standalone meaning and does not occur outside of a verbal MWE. For example, *tashkil* in the MWE *tashkil qil* ‘establish’ does not belong to any POS in Uzbek and the whole phrase is considered a verb in traditional Uzbek grammar. However, UD requires to analyze this phrase as two tokens tagged `PART` and `VERB`, respectively. The main challenge is the lack of a comprehensive list of such MWEs, requiring frequent dictionary lookups to verify if the first element of the verb phrase belongs to a different POS category.

With regard to Universal Features, Uzbek verbs can be morphologically marked for the Voice category by more than one value. In such cases, the actual value is determined by the most external voice suffix. For instance, *ko’ch-ir-il-ish-i* ‘relocate-CAU-PASS-VNOUN-3SG’ has a causative and a passive morpheme, but the verb is ultimately considered as having a passive voice. This ambiguity should be resolved manually, as the parser has no representation for the order of the morphemes.

As for syntactic relations, postverbial constructions with auxiliary verbs, which are defined by Johanson (2021, 36-37) as “converb[s] of a lexical verb and a second auxiliary verb form[ing] a verbal phrase with strong semantic fusion”, are notoriously challenging to analyze. There are about 27 verbs in Uzbek that can be used as auxiliaries to form such constructions, e.g. *to‘xtab* ‘stop’ as in *to‘xtab qol* (‘lit.: stopping stay) ’end, finish’ (see Figure 2) (Boeschoten, 2021b, 396).

Postverbial constructions are common in the Turkic family, but their annotation lacks consistency across the UD treebanks for Turkic languages. In the Uyghur treebank, auxiliaries are analyzed as the head of an open clausal complement relation (`xcomp`)[10], although this does not fully align with the UD guidelines. In the Kyrgyz treebank, converbs are treated as the head of the relation, with the postverbial element assigned an auxiliary relation (`aux`)[11]. However, this seems inaccurate, as verbal features like person, tense and mood are marked on the postverbial element. In Uzbek, words used as auxiliaries also have non-auxiliary uses, and `aux` is only assigned to modal and copular verbs. This inconsistency across languages highlights the need for a standardized approach. One potential solution is to introduce a new subtype for compound relations, pending discussion among Turkic UD contributors and approval by the UD coordinators. In the meantime, we analyze such Uzbek verb constructions with a `compound` relation, in which the postverbial element serves as the head.

## 6 Conclusion

In this work, we presented the first UD treebank for Uzbek – Uzbek-UT. The annotation methodology was semi-automatic, starting from manual annotation of training data to automatic parsing with freely available tools, followed by human post-editing. Additionally, we analyzed constructions that are particularly challenging in the UD framework. Despite its small size, the treebank serves

[10] `https://universaldependencies.org/ug/index.html`

[11] `https://github.com/UniversalDependencies/UD_Kyrgyz-TueCL`

as a quality resource for linguistic research and model training in several NLP tasks, which we intend to conduct in future work. In the future, this treebank can be extended in size, covering more registers and enriched with additional tags and improved solutions for MWEs.

## References


Alessandro Agostini, Timur Usmanov, Ulugbek Khamdamov, Nilufar Abdurakhmonova, and Mukhammadsaid Mamasaidov. 2021. UZWORDNET: A lexical-semantic database for the Uzbek language. In *Proceedings of the 11th Global Wordnet Conference*, pages 8–19, University of South Africa (UNISA). Global Wordnet Association.

Hendrik Boeschoten. 2021a. The speakers of Turkic languages. In Lars Johanson and Éva Á. Csató, editors, *The Turkic languages*, pages 1–12. Routledge.

Hendrik Boeschoten. 2021b. Uzbek. In Lars Johanson and Éva Á. Csató, editors, *The Turkic languages*, pages 388–408. Routledge.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175, Online. Association for Computational Linguistics.

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

Elmurod Kuriyozov, Sanatbek Matlatipov, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2019. Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek. In *Human Language Technology. Challenges for Computer Science and Linguistics: 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17–19, 2019, Revised Selected Papers*, page 232–243, Berlin, Heidelberg. Springer-Verlag.

Elmurod Kuriyozov, Ulugbek Salaev, Sanatbek Matlatipov, and Gayrat Matlatipov. 2023. Text classification dataset and analysis for Uzbek language. *CoRR*, abs/2302.14494.

Elmurod Kuriyozov, David Vilares, and Carlos Gómez-Rodríguez. 2024. BERTbek: A pretrained language model for Uzbek. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 33–44, Torino, Italia. ELRA and ICCL.

Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*, volume 1, page 63–70. Association for Computational Linguistics.

Kh. A. Madatov, D. J. Khujamov, and B. R. Boltayev. 2022. Creating of the Uzbek WordNet based on Turkish WordNet. In *AIP Conference Proceedings*, volume 2432. AIP Publishing.

Mukhammadsaid Mamasaidov and Abror Shopulatov. 2023. Uzcrawl dataset.

B. Mansurov and A. Mansurov. 2021. UzBERT: pretraining a BERT model for Uzbek. *CoRR*, abs/2108.09814.

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

Gayrat Matlatipov and Zygmunt Vetulani. 2009. *Representation of Uzbek Morphology in Prolog*, page 83–110. Springer-Verlag, Berlin, Heidelberg.

Sanatbek Matlatipov, Hulkar Rahimboeva, Jaloliddin Rajabov, and Elmurod Kuriyozov. 2022. Uzbek sentiment analysis based on local restaurant reviews. In *Proceedings of the ALTNLP The International Conference and workshop on Agglutanative Language Technologies as a challenge of Natural Language Processing, Virtual Event, Koper, Slovenia, June, 7th and 8th, 2022*, volume 3315 of *CEUR Workshop Proceedings*, pages 126–136. CEUR-WS.org.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

I. M. Rabbimov and S. S. Kobilov. 2020. Multi-class text classification of Uzbek news articles using machine learning. *Journal of Physics: Conference Series*, 1546(1):012097.

Shavkat Rahmatullayev. 2006. *Hozirgi Adabiy O'zbek Tili [Contemporary Literary Uzbek Language (textbook)]*. Universitet.

Ulugbek Salaev. 2023. Modeling morphological analysis based on word-ending for Uzbek language. *Science and innovation*, 2(C11):29–34.

Ulugbek Salaev, Elmurod Kuriyozov, and Carlos Gómez-Rodríguez. 2022. SimRelUz: Similarity and relatedness scores as a semantic evaluation dataset for Uzbek language. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 199–206, Marseille, France. European Language Resources Association.

Maksud Sharipov, Elmurod Kuriyozov, Ollabergan Yuldashev, and Ogabek Sobirov. 2023. Uzbektagger: The rule-based POS tagger for Uzbek language. *arXiv preprint arXiv:2301.12711*.

Maksud Sharipov, Jamolbek Mattiev, Jasur Sobirov, and Rustam Baltayev. 2022. Creating a morphological and syntactic tagged corpus for the Uzbek language. In *Proceedings of the ALTNLP The International Conference and workshop on Agglutanative Language Technologies as a challenge of Natural Language Processing, Virtual Event, Koper, Slovenia, June, 7th and 8th, 2022*, volume 3315 of *CEUR Workshop Proceedings*, pages 93–98. CEUR-WS.org.

Maksud Sharipov and Ogabek Sobirov. 2022. Development of a rule-based lemmatization algorithm through finite state machine for Uzbek language. In *Proceedings of the ALTNLP The International Conference and workshop on Agglutanative Language Technologies as a challenge of Natural Language Processing, Virtual Event, Koper, Slovenia, June, 7th and 8th, 2022*, volume 3315 of *CEUR Workshop Proceedings*, pages 154–159. CEUR-WS.org.

Maksud Sharipov and Ollabergan Yuldashov. 2022. Uzbekstemmer: Development of a rule-based stemming algorithm for Uzbek language. In *Proceedings of the ALTNLP The International Conference and workshop on Agglutanative Language Technologies as a challenge of Natural Language Processing, Virtual Event, Koper, Slovenia, June, 7th and 8th, 2022*, volume 3315 of *CEUR Workshop Proceedings*, pages 137–144. CEUR-WS.org.