UzUDT: Uzbek Universal Dependencies Treebank

Sanatbek Matlatipov[1], **Mersaid Aripov**[1]
[1]National University of Uzbekistan named after Mirzo Ulugbek
University street, 4th house, Tashkent, Uzbekistan
{s.matlatipov, mirsaid.aripov}@nuu.uz

Abstract

In this paper, we present a new Universal Dependencies treebank for Uzbek language(UzUDT) developed as a gold-standard resource with full manual annotation. The treebank includes 684 sentences (7,582 tokens) from Uzbek literary texts, and is larger and more domain-diverse than the existing Uzbek UD treebank. The corpus was developed through rigorous multi-annotator adjudication, achieving very high inter-annotator agreement (multi-rater agreement coefficients > 0.90) across lemmatization, POS tagging, and morphological features. Alongside comprehensive corpus profiling, we establish robust computational baselines by evaluating graph-based (Stanza) and transition-based (spaCy) parsing architectures using both static and monolingual contextual embeddings. Our evaluations reveal a critical architectural trade-off for low-resource agglutinative parsing: joint transition-based models excel at morphosyntactic tagging, whereas graph-based models remain strictly superior for resolving complex structural dependencies. Furthermore, we demonstrate that cross-treebank data augmentation yields substantial, synergistic accuracy gains, achieving a peak Universal Part-of-Speech (UPOS) tagging accuracy of 89.18%, a Labeled Attachment Score (LAS) of 63.81%, and a morphological feature accuracy of 71.09%. The resource provides much-needed high-quality treebank for Uzbek to assist in developing better NLP tools and to enable linguistic research in the low-resource language.

Keywords: Universal Dependencies; dependency parsing; Uzbek

1. Introduction

Uzbek is a low-resource Turkic language and had very few core NLP resources until recently(Veitsman and Hartmann, 2025). In particular, dependency treebanks were rare in Uzbek. The first Uzbek UD treebank, UD_Uzbek-UT(Akhundjanova and Talamo, 2025), was released recently and provides 500 sentences (5,930 tokens) from news and fiction domains, establishing an important initial foundation for Uzbek dependency parsing. The first Uzbek UD treebank (Uzbek-UT) (Akhundjanova and Talamo, 2025) contains 500 sentences (5,930 tokens) from news and fiction domains and constitutes an important initial contribution to Uzbek dependency parsing. Apart from that preliminary attempt, Uzbek had very little computational resources: an old morphological parser was built in Prolog(Matlatipov and Vetulani, 2009) and rule-based tools (e.g., UzbekTagger by (Sharipov et al., 2023, 2024)) and a morphological analyzer (UzMorphAnalyzer by (Salaev, 2024), (Abdurakhmonova et al., 2022), (Matlatipov et al., 2020)) have been recently created. However, publicly available Uzbek UD resources remain small and genre-limited, which constrains robust tagger/parser training and cross-domain evaluation. To cover the resource gap, we have created a gold-standard UD treebank for Uzbek larger in size and with a different domain scope than the current Uzbek-UT treebank (Akhundjanova and Talamo, 2025). Our treebank contains 684 sentences; this larger and more diverse treebank can improve the robustness of Uzbek language models and parsers and better support downstream applications (e.g., literary NLP tools, information extraction, and educational technologies).

In summary, we have made the following contributions:

- **New Uzbek UD Treebank:** We present a new Uzbek Universal Dependencies (UD) treebank containing 684 sentences (approximately 7.6k tokens) from literature texts, manually annotated with full morphosyntactic structure. To our knowledge, UzUDT is currently the largest publicly released Uzbek UD treebank by sentence count, and it complements Uzbek-UT by covering a different genre.
- **High-Quality Annotation Process:** The annotation was carried out by six annotators with rigorous agreement checks, achieving extremely high agreement between annotators (multi-rater agreement coefficients > 0.90; Table 1) on core annotations. We document our annotation process using the INCEpTION tool(Klie et al., 2018) and explain how disagreements were resolved to produce a gold-standard dataset.
- **Comprehensive Data Analysis:** We conduct a detailed analysis of the treebank, including token statistics, vocabulary statistics, frequency statistics of POS tags, and frequencies of morphological features and dependency relations. We compare these statistics with those from the previous Uzbek UD treebank to highlight differences resulting from the larger dataset.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11642–11649
11–16 May 2026.

- **Parse Evaluation Tests:** To establish robust baselines for low-resource agglutinative parsing, we evaluate and compare two distinct architectural paradigms: a graph-based pipeline (Stanza(Qi et al., 2020)) and a transition-based pipeline (spaCy(Honnibal and Montani, 2017)). Furthermore, we analyze the impact of representation learning by comparing static embeddings (FastText(Bojanowski et al., 2017)) against monolingual contextual embeddings (TahrirchiBERTMamasaidov and Shopulatov (2023)) across both standalone and merged treebank configurations. We report comprehensive metrics including Labeled Attachment Score (LAS), Unlabeled Attachment Score (UAS), and morphosyntactic tagging accuracy. Our findings demonstrate that while joint transition-based models excel at tagging, graph-based models are vastly superior for structural parsing, and cross-treebank augmentation yields substantial accuracy gains. The training and testing code, along with guidelines, are publicly available in our GitHub repository[1].

Our hope is that the new treebank will stimulate further Uzbek NLP work. It provides a much-needed resource to construct better taggers and parsers and can serve as a foundation for multilingual and cross-lingual work with Uzbek. We release the treebank under an open license to contribute to the growing global collection of Universal Dependencies resources[2]. UzUDT treebank is distributed under the Creative Commons Attribution–ShareAlike 4.0 license (CC BY-SA 4.0).

2. Methodology

2.1. Data collection

We compiled UzUDT by selecting isolated sentences from the works *Kun shundan boshlanadi* and *Maqar*. These sentences are included under an explicit non-exclusive permission granted by the author and copyright holder, allowing sentence-level selection, UD-style linguistic (de Marneffe et al., 2021) annotation, model training/evaluation, and open distribution of the selected sentences as part of the UzUDT release (see 5). We restricted selection to self-contained sentences in the Uzbek Latin script, avoiding heavily dialectal or non-standard forms. The resulting corpus contains 684 sentences (7,582 tokens) with an average length of approximately 11 tokens per sentence. We split the data into 451/45/188 sentences for train/dev/test, respectively. Figure 1 illustrates a complete dependency graph extracted from our dataset, exhibiting the rich morphological and syntactic UD annotations required for complex Uzbek constructions, such as serial verbs (`compound:svc`) and chained modifiers.

2.2. Annotation Process

The annotation followed the Universal Dependencies (UD) guidelines (de Marneffe et al., 2021) for tokenization, lemmatization, POS tagging, morphological features, and dependency relations. We followed the Uzbek-specific UD documentation and contributed to maintaining the Uzbek UD feature inventory (e.g., PronType; `https://universaldependencies.org/uz/feat/PronType.html`). We pre-segmented sentences into tokens according to UD rules, so annotators did not perform tokenization. Annotation was carried out in the INCEpTION platform (Klie et al., 2018) with custom layers aligned to UD fields. The team consisted of six annotators (four Uzbek linguists and two NLP engineers), who completed an initial training phase by annotating a shared calibration subset and discussing Uzbek-specific UD cases (e.g., agglutinative case marking and null copulas).

Inter-annotator agreement (IAA)(Richie et al., 2022) was measured on the shared calibration subset at the token level for (i) lemma, (ii) UPOS, and (iii) complete morphological feature bundles (UFeats, evaluated as exact-match labels per token). We report multi-rater agreement using Fleiss' κ(Wang and Yusof, 2025; Fleiss, 1971) and Krippendorff's α(Ángel González-Prieto et al., 2023) (nominal), observing very high agreement across the three layers (Table 1). We treat all labels as nominal and compute agreement at the token level; for UFeats we require an exact match of the complete feature bundle per token. The full dataset was then annotated in a double-annotation setting (each sentence independently annotated by two annotators). Disagreements were identified using INCEpTION's comparison view and resolved in regular adjudication meetings, with a senior linguist acting as the final arbiter in difficult cases. The resulting treebank is a single gold-standard annotation per sentence, exported in CoNLL-U format.

2.3. Data Analysis

Following the completion of the annotation process, we conducted a rigorous linguistic and statistical diagnostic of UzUDT using the official Universal Dependencies validation toolkit. To contextualize our contribution, we contrast UzUDT against the existing Uzbek-UT treebank, focusing on how domain

[1] https://github.com/SanatbekMatlatipov/uzudtevaluations

[2] https://github.com/UniversalDependencies/UD_Uzbek-UzUDT/tree/dev

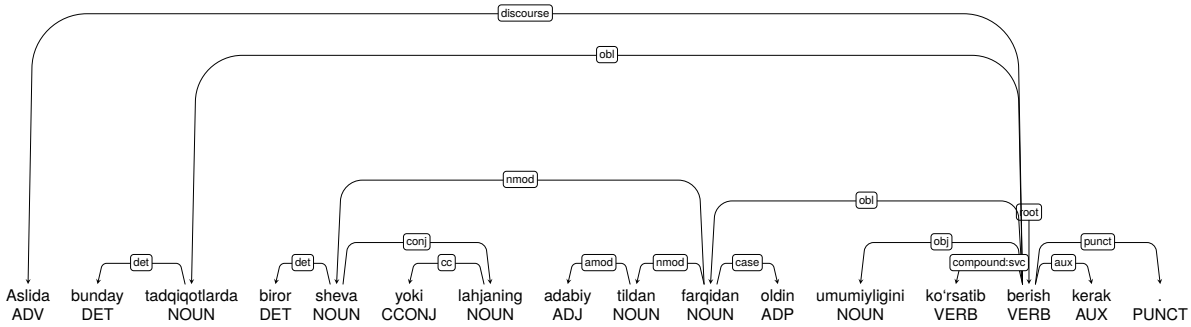

Figure 1: A complete syntactic dependency tree (UD annotation) extracted from the UzUDT corpus. This example illustrates the rich morphology characteristic of the Uzbek language, specifically highlighting chained case markers and serial verb (`compound:svc`) constructions.

Annotation layer	Fleiss' κ	Krippendorff's α
Lemma	0.952	0.946
UPOS	0.937	0.941
UFeats	0.908	0.899

Table 1: Multi-rater inter-annotator agreement. Scores are token-level agreement coefficients (nominal)

divergence (literary vs. news/fiction) influences morphological and syntactic distributions. UzUDT passes the official UD validation checks without errors, indicating a consistent CoNLL-U release compliant with the UD guidelines.

Corpus Statistics and Lexical Diversity Table 2 summarizes the high-level linguistic statistics. UzUDT contains 684 sentences and 7,582 surface tokens (yielding an average of 11.1 tokens per sentence). While UzUDT provides a larger absolute volume of tokens than the previous treebank, it exhibits a lower Type-Token Ratio (TTR)(Richards, 1987) of 0.409 compared to UT's 0.571. This lower lexical sparsity is characteristic of narrative prose, which heavily reuse core vocabulary, whereas the news domain in Uzbek-UT inherently introduces a broader spectrum of unique named entities.

Part-of-Speech Distribution The universal part-of-speech (UPOS) distributions strongly reflect the underlying text genres. Nouns (33.3%) and verbs (20.9%) remain the dominant categories in UzUDT. However, the literary nature of our data yields a significantly higher proportion of pronouns (6.0% vs. 3.3%). Conversely, proper nouns (`PROPN`) are extremely sparse in our dataset (0.3%) compared to the news-heavy Uzbek-UT corpus (5.2%). A comprehensive UPOS breakdown is provided in Appendix A.

Linguistic Metric	UzUDT	Uzbek-UT
Corpus Size		
Sentences	684	500
Tokens (Surface)	7,582	5,930
Type-Token Ratio (Forms)	0.409	0.571
Avg. Tokens / Sentence	11.1	11.9
Key POS Proportions (%)		
PRON (Pronouns)	6.0%	3.3%
PROPN (Proper Nouns)	0.3%	5.2%
Key Syntactic Relations (%)		
`advcl` (Adverbial Clause)	5.9%	1.8%
`compound:lvc` (Light Verb)	0.3%	3.6%

Table 2: Comparative linguistic statistics highlighting domain-driven differences between the new UzUDT (Literary) and the previous Uzbek-UT (News/Fiction) treebanks.

Morphological Richness Uzbek is a highly agglutinative language, and UzUDT captures this complexity with a richer morphosyntactic inventory than Uzbek-UT. Notably, our annotation explicitly captures possessor agreement on nouns (NUMBER[PSOR] and PERSON[PSOR]) and distinguishes firsthand versus non-firsthand evidentiality (EVIDENT=FH/NFH)—a typologically crucial Turkic feature that was previously unannotated. Furthermore, UzUDT provides finer granularity in verbal aspect, capturing HABITUAL, IMPERFECTIVE, and PERFECTIVE distinctions alongside the standard PROGRESSIVE form. The full morphological inventory is documented in Appendix B.

Syntactic Dependencies While core arguments (`nsubj`, `obj`, `obl`) maintain stable frequencies across both datasets, mid-frequency relations expose structural disparities. Literary Uzbek relies heavily on complex converb chains (using suffixes like *-ib* or *-gach*) to link sequential actions. Consequently, adverbial clauses (`advcl`) are more than three times as frequent in UzUDT (5.9%) as in the

news corpus (1.8%). Figure 2 illustrates a typical converb construction extracted from our test set. In contrast, the news domain relies heavily on light-verb constructions (`compound:lvc`), which appear at ten times the frequency in UT. These complementary distributions underscore the necessity of merging these treebanks to train robust, domain-agnostic parsers. The complete comparative dependency inventory is detailed in Appendix C.

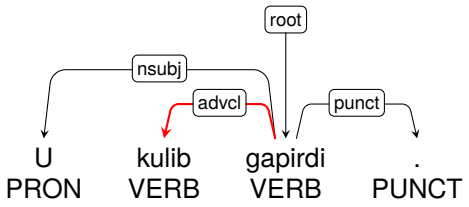

Figure 2: A simplified dependency tree demonstrating an adverbial clause (`advcl`) driven by a converb (*kulib* "smilingly"). This syntactic structure is highly prevalent in the UzUDT literary texts.

Cross-Treebank Annotation Divergence While data augmentation significantly improves overall parsing accuracy, merging heterogeneous treebanks introduces annotation friction. Our post-evaluation error analysis reveals that morphological features unique to a single treebank's guidelines suffer from severe recall degradation when trained on the merged dataset. For example, possessor agreement (NUMBER[PSOR]) and evidentiality (EVIDENT) are explicitly annotated in UzUDT but absent in UT. Consequently, the merged transition-based model yields a low recall for NUMBER[PSOR] (43.7%) and entirely fails to predict EVIDENT (0.0%). The model struggles to reconcile the explicit presence of these features in UzUDT with their systematic absence in UT, highlighting that future cross-treebank augmentation in low-resource settings requires rigorous harmonization of morphological guidelines.

3. Experiments

To assess the usability of the new treebank and establish robust baselines, we evaluated morphosyntactic tagging and dependency parsing under varying low-resource constraints. We compared representation strategies (static vs. contextual) and architectural paradigms (graph-based vs. transition-based) strictly on UzUDT, and on a merged dataset of UzUDT and UD_Uzbek-UT.

Experimental Setup: Our primary pipeline utilizes a modified Stanza framework featuring a shared BiLSTM(Kiperwasser and Goldberg, 2016) encoder. Tagging (UPOS/XPOS/UFeats) is handled by a softmax multi-task head, while syntax is predicted via a DeepBiaffine graph-based parser. We trained this architecture using Adam ($\beta = (0.9, 0.999)$), a learning rate of 3×10^{-3}, a batch size of 5,000 tokens, and a dropout of 0.33, comparing static FastText embeddings against contextual TahrirchiBERT subwords. We obtain token-level contextual embeddings from TahrirchiBERT by aligning subwords to UD tokens and applying a fixed pooling rule (last-subword or mean pooling); these token representations are then used as inputs to Stanza's tagger and biaffine parser across all contextual experiments. The training dynamics and convergence trajectories for this Stanza-based joint morphosyntactic tagger and dependency parser are illustrated in **Figure 3** and **Figure 4**, respectively.

To provide a contrasting baseline, we additionally trained a joint, end-to-end transition-based pipeline (spaCy) featuring an arc-eager parser and the identical TahrirchiBERT encoder. The spaCy model was optimized using Adam with an initial learning rate of 5×10^{-5} (linear decay) and a compounding batch size (100–1,000 words). Both architectures utilized early stopping evaluated on the development set.

Reproducibility. All experiments were run with fixed random seeds $\{13, 21, 42\}$; we report the run achieving the best development-set score, and select the final checkpoint by development-set performance. Training and evaluation were run on a single NVIDIA RTX A6000 GPU (48 GB VRAM) with CUDA 12.4. We use Stanza v1.4.0, spaCy v3.8.11, PyTorch v2.6.0+cu124, and Transformers v4.49.0. Complete environment details and run configurations are provided in the accompanying repository and Weights & Biases logs.

Results: Table 3 details the Universal Part-of-Speech (UPOS), Morphological Features (UFeats), Unlabeled Attachment Score (UAS), and Labeled Attachment Score (LAS) across the test sets.

The results reveal three critical findings for low-resource agglutinative parsing. First, replacing static FastText(Bojanowski et al., 2017; Joulin et al., 2016) embeddings with TahrirchiBERT yields consistent improvements across metrics. Second, cross-treebank data augmentation provides the most substantial boost to structural parsing; expanding the training data improved the Stanza graph-based LAS by roughly 10 to 11 points.

Finally, comparing architectural paradigms reveals a distinct trade-off. The joint transition-based pipeline (spaCy) excels at basic tagging, achieving the highest UPOS (89.18%) on the merged data.

Parser	Emb.	Data	UPOS	UFeat	UAS	LAS
Stanza	FT	UzUDT	79.19	66.61	69.57	51.24
Stanza	FT	Merged	80.26	66.98	72.27	62.40
Stanza	TB	UzUDT	82.45	65.37	72.05	54.19
Stanza	TB	Merged	85.08	**71.09**	**72.39**	**63.81**
spaCy	TB	UzUDT	86.50	50.55	67.72	45.35
spaCy	TB	Merged	**89.18**	65.48	66.81	47.11

Table 3: Parsing and tagging performance comparing architectures and representations. Abbreviations – Emb: FT (FastText), TB (TahrirchiBERT). Data: Merged (UzUDT + Uzbek-UT). Stanza utilizes a graph-based parser, while spaCy utilizes a transition-based parser.

This success is driven by its end-to-end objective, which reinforces tagging signals through the parser loss; furthermore, augmented data triggered a massive +14.93 point jump in its morphological feature (UFeats) accuracy. In contrast, the graph-based parser (Stanza) vastly outperforms the transition-based approach on labeled structural parsing (LAS 63.81% vs. 47.11%). This 16.70-point advantage demonstrates that graph-based global decoding is better equipped to resolve complex, long-distance dependencies in head-final (SOV) Uzbek syntax.(Aripov et al., 2022)

4. Conclusion

We have published UzUDT, a new gold-standard Universal Dependencies treebank for the Uzbek language containing 684 sentences from literary texts. It significantly enhances the availability of Uzbek annotated data in terms of domain variety and size. with a rigorous annotation process executed through the INCEpTION platform, we achieved very high multi-rater agreement (Fleiss' κ and Krippendorff's $\alpha > 0.90$) for lemma, UPOS, and full morphological feature bundles. Our analysis demonstrates that UzUDT provides comprehensive coverage of Uzbek's rich agglutinative morphology and complex syntactic structures, such as converb-driven adverbial clauses. To establish robust computational baselines, we evaluated multiple parsing paradigms. Our findings reveal a critical architectural trade-off for low-resource agglutinative languages: while joint transition-based pipelines (spaCy) excel at morphosyntactic tagging by leveraging end-to-end objective reinforcement, graph-based parsers (Stanza) remain vastly superior for resolving complex, long-distance structural dependencies (LAS). Also, we demonstrated that merging heterogeneous treebanks and upgrading to monolingual contextual representations yields substantial accuracy gains, provided that cross-treebank annotation divergences are carefully managed.

Future Work. We will scale the resource beyond literary text by expanding to additional genres (e.g., news, web text, and social media) using model-assisted annotation: our strongest graph-based parser will be used to pre-annotate new data, with annotators focusing on correction and adjudication. Moreover, we will investigate LLM-assisted syntactic annotation(Schroeder et al., 2025) as an alternative scaling strategy, comparing (i) conventional neural parsers trained on UD data and (ii) instruction-tuned LLM pipelines that generate UD-style analyses and can be refined with human feedback. Quality control will combine automatic UD validation with LLM-as-a-judge(Hu et al., 2025) scoring and selective human review to measure accuracy, consistency, and annotation cost at scale. Ultimately, our goal is to enable robust, publicly available Uzbek models in widely used NLP toolkits, including Stanza and spaCy.

5. Ethics Statement

UzUDT is released as an open research resource for Uzbek NLP with attention to copyright and responsible data use. The corpus includes sentence-level material from the literary works *Kun shundan boshlanadi* and *Maqar*. We obtained signed written permission from the author and copyright holder to select sentences from these works, annotate them with UD morphology and syntax, use them for model training and evaluation, and distribute the selected sentences as part of the UzUDT release with appropriate attribution. A copy of the permission agreement can be provided to the organizers and reviewers upon request.

6. Acknowledgements

The authors gratefully acknowledge Elmurod Kuriyozov, Komilova Madinabonu, and Matyoqubova Shaydo for their assistance with dataset annotation.

7. Bibliographical References

Nilufar Abdurakhmonova, Ismailov Alisher, and Rano Sayfulleyeva. 2022. Morphuz: Morphological analyzer for the uzbek language. In *2022 7th International Conference on Computer Science and Engineering (UBMK)*, pages 61–66.

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. Association for Computational Linguistics.

Mersaid Aripov, Muftakh Khakimov, Sanatbek Matlatipov, and Ziyoviddin Sirojiddinov. 2022. Analysis and processing of the uzbek language on the multi-language modelled computer translator technology. In *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 81–95, Cham. Springer International Publishing.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

M Honnibal and I Montani. 2017. *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Renjun Hu, Yi Cheng, Libin Meng, Jiaxin Xia, Yi Zong, Xing Shi, and Wei Lin. 2025. Training an llm-as-a-judge model: Pipeline, insights, and practical lessons. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 228–237, New York, NY, USA. Association for Computing Machinery.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

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

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*, pages 101–108, Online. Association for Computational Linguistics.

Brian Richards. 1987. Type/token ratios: what do they really tell us? *Journal of Child Language*, 14(2):201–209.

Russell Richie, Sachin Grover, and Fuchiang (Rich) Tsui. 2022. Inter-annotator agreement is not the ceiling of machine learning performance: Evidence from a comprehensive set of simulations. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 275–284, Dublin, Ireland. Association for Computational Linguistics.

Ulugbek Salaev. 2024. Uzmorphanalyser: A morphological analysis model for the uzbek language using inflectional endings. *AIP Conference Proceedings*, 3244(1):030058.

Hope Schroeder, Deb Roy, and Jad Kabbara. 2025. Just put a human in the loop? investigating LLM-assisted annotation for subjective tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25771–25795, Vienna, Austria. Association for Computational Linguistics.

Maksud Sharipov, Elmurod Kuriyozov, Ollabergan Yuldashev, and Ogabek Sobirov. 2023. Uzbektagger: The rule-based pos tagger for uzbek language.

Maksud S. Sharipov, Hushnudbek S. Adinaev, and Elmurod R. Kuriyozov. 2024. Rule-based punctuation algorithm for the uzbek language. In *2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM)*, pages 2410–2414.

Yana Veitsman and Mareike Hartmann. 2025. Recent advancements and challenges of Turkic Central Asian language processing. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 309–324, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ying Wang and Ibnatul Jalilah Yusof. 2025. Expert consensus and reliability validation of the

portfolio assessment guideline for chinese practical writing: An empirical study based on fleiss' kappa. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 5(4):100248.

Ángel González-Prieto, Jorge Perez, Jessica Diaz, and Daniel López-Fernández. 2023. Reliability in software engineering qualitative research through inter-coder agreement. *Journal of Systems and Software*, 202:111707.

8. Language Resource References

Mukhammadsaid Mamasaidov and Abror Shopulatov. 2023. *TahrirchiBERT base*. Accessed: 2026-02-28.

Sanatbek Matlatipov, Ualsher Tukeyev, and Mersaid Aripov. 2020. Towards the uzbek language endings as a language resource. In *Advances in Computational Collective Intelligence*, pages 729–740, Cham. Springer International Publishing.

A. Comprehensive POS Distribution

Table 4 provides the complete distributional comparison of UPOS tags between UzUDT and the previous UD_Uzbek-UT treebank.

	UzUDT (Ours)		Uzbek-UT	
UPOS	**Count**	**%**	**Count**	**%**
NOUN	2,526	33.3	2,152	36.3
VERB	1,585	20.9	987	16.6
PUNCT	1,571	20.7	860	14.5
ADJ	522	6.9	484	8.2
PRON	458	6.0	193	3.3
PROPN	26	0.3	308	5.2
ADV	230	3.0	203	3.4
NUM	184	2.4	214	3.6
ADP	120	1.6	184	3.1
DET	103	1.4	103	1.7
CCONJ	87	1.1	104	1.8
AUX	88	1.2	79	1.3
PART	43	0.6	36	0.6
INTJ	23	0.3	6	0.1
SCONJ	9	0.1	9	0.2
X	7	0.1	7	0.1
SYM	0	0.0	1	0.0

Table 4: Full UPOS distribution comparing the two Uzbek UD treebanks.

B. Morphological Feature Inventory

Morphology with unique features.

Category	**Values in UzUDT**	**Status vs. Uzbek-UT**
Aspect	Hab, Imp, Perf, Prog	UzUDT has 3 unique
Case	Abl, Acc, Dat, Gen, Loc, Nom	Identical
Evident	Fh, Nfh	**Unique to UzUDT**
Mood	Cnd, Des, Imp, Ind, Int, Opt, Pot	UzUDT has *Des*
Number	Plur, Sing	Identical
Number[psor]	Plur; Plur,Sing; Sing	**Unique to UzUDT**
Person	1, 2, 3	Identical
Person[psor]	1, 2, 3	**Unique to UzUDT**
Tense	Fut, Past, Pres	Identical
VerbForm	Conv, Fin, Inf, Part, Vnoun	Identical
Voice	Act, Pass	UzUDT has *Act*
PronType	Dem, Ind, Int, Neg, Prs, Rcp, Rel, Tot	Identical
Polarity	Neg, Pos	UzUDT has *Pos*
Reflex	Yes	Identical
Poss	Yes	Identical

Table 5: Primary morphological feature categories present in UzUDT.

C. Dependency Relation Inventory

	UzUDT (Ours)		Uzbek-UT	
Relation	**Count**	**%**	**Count**	**%**
punct	1,571	20.7	860	14.5
obl	720	9.5	602	10.1
nsubj	720	9.5	540	9.1
amod	452	6.0	338	5.7
advcl	444	5.9	106	1.8
obj	418	5.5	251	4.2
compound	265	3.5	290	4.9
nmod	252	3.3	350	5.9
advmod	216	2.8	238	4.0
conj	200	2.6	208	3.5
nmod:poss	189	2.5	142	2.4
det	183	2.4	103	1.7
xcomp	139	1.8	-	-
acl	137	1.8	133	2.2
nummod	131	1.7	160	2.7
dep	125	1.6	0	0.0
cc	109	1.4	109	1.8
ccomp	94	1.2	-	-
compound:lvc	21	0.3	215	3.6
flat	11	0.1	106	1.8

Table 6: Top syntactic dependency relations highlighting structural differences.

Figure 3: Training dynamics for the joint morphosyntactic tagger.

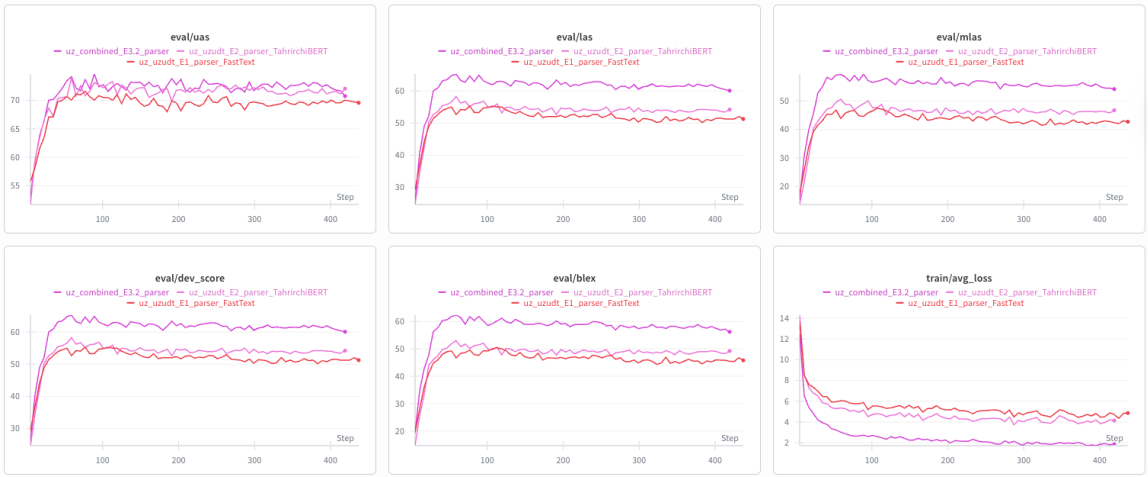

Figure 4: Training dynamics for the dependency parser.