CORPUS LINGUISTICS IN UZBEKISTAN: CURRENT STATE, CHALLENGES, AND PROSPECTS FOR UZBEK LANGUAGE RESEARCH

Gulnoraxon Toshtemirova

Kokand University PhD candidate

E-mail: gulnoraxontoshtemirova2909@gmail.com

ABSTRACT

This article investigates the development and current status of corpus linguistics in Uzbekistan, focusing on the construction, availability, and application of Uzbek-language corpora in linguistic research and language education. Using a mixed-methods approach that combines systematic review of existing corpora and published research with semi-structured interviews with ten Uzbek corpus linguists, the study identifies critical infrastructure gaps, methodological challenges, and institutional barriers that constrain the field. Results reveal that while several Uzbek corpora exist — including the Uzbek National Corpus and domain-specific sub-corpora — they remain significantly smaller and less annotated than comparable corpora for major world languages. Furthermore, corpus-based methods are underutilised in Uzbek linguistics pedagogy and applied language research. The article concludes with evidence-based recommendations for scaling corpus infrastructure, fostering interdisciplinary collaboration, and integrating corpus tools into university curricula. The findings contribute to the growing literature on corpus linguistics in under-resourced language contexts and provide a roadmap for Uzbek language technology development.

Keywords: corpus linguistics, Uzbek language, language corpus, NLP, under-resourced languages, Uzbekistan

INSTRODUCTION

Corpus linguistics has transformed the empirical study of language over the past six decades, providing researchers with large, machine-readable collections of authentic texts

that enable statistically reliable observations about lexis, grammar, discourse, and language change (McEnery & Hardie, 2012; Biber, Conrad & Reppen, 1998). Pioneering resources such as the Brown Corpus (1961), the British National Corpus (1994), and the Corpus of Contemporary American English (COCA) have driven paradigm shifts in descriptive linguistics, lexicography, and language teaching. However, this progress has been distributed unevenly: languages with large speaker populations and strong institutional backing enjoy rich corpus resources, while many other languages — particularly those of Central Asia — remain under-resourced (Choudhury & Jha, 2015). Uzbek, the official state language of the Republic of Uzbekistan and a member of the Turkic language family, is spoken by approximately 35–37 million people worldwide (Ethnologue, 2023). Despite this considerable speaker base, the development of large-scale, publicly accessible corpus resources for Uzbek has lagged behind comparable Turkic languages such as Turkish (Turkish National Corpus) and Kazakh (Kazakh National Corpus). This disparity has practical consequences: Uzbek natural language processing (NLP) systems, machine translation tools, and pedagogical grammars all suffer from inadequate empirical foundations. The post-independence period since 1991 has brought significant changes to the linguistic landscape of Uzbekistan. The transition from the Cyrillic to a Latin-based script — formalised in 1993 and still ongoing in public life — has created unique challenges for corpus compilation, as texts exist in multiple orthographic conventions (Sjoberg, 1963; Johanson, 1998). Simultaneously, the government's language policy has elevated the status of Uzbek in education, administration, and media, generating new registers and genres of written Uzbek that warrant systematic corpus-based investigation (Landau & Kellner-Heinkele, 2001).

This study addresses the following research questions: (1) What corpus resources currently exist for the Uzbek language, and what are their principal characteristics in terms of size, annotation, and accessibility? (2) What methodological and institutional challenges do Uzbek corpus linguists encounter? (3) How are corpus methods currently integrated

into linguistic research and language education in Uzbekistan? (4) What steps are necessary to advance corpus linguistics as a discipline in Uzbekistan?

The article is structured according to the IMRaD format. Section 2 reviews the relevant literature on corpus linguistics in under-resourced language contexts and the specific situation of Uzbek. Section 3 describes the mixed-methods research design. Section 4 presents quantitative and qualitative findings. Section 5 discusses implications and provides recommendations, followed by a conclusion.

Corpus linguistics is defined as the study of language based on examples of real-life language use (McEnery & Wilson, 2001, p. 1). Since Sinclair's (1991) foundational work on lexis and the idiom principle, the field has expanded to encompass learner corpora (Granger, 2003), historical corpora (Rissanen, 1994), spoken corpora (Carter & McCarthy, 2017), and parallel corpora for translation studies (Baker, 1995). Key methodological tenets include representativeness, authenticity, and machine readability. A recurrent finding in the literature is that corpus size, genre balance, and annotation depth critically determine the types of linguistic claims that researchers can validly make (Biber, 1993; Kilgarriff & Grefenstette, 2003).

The concept of language resourcing — the degree to which a language is supported by digital tools, corpora, and NLP systems — has attracted growing scholarly attention. Choudhury and Jha (2015) propose a taxonomy of 'low-resource' languages based on data availability, speaker population, and institutional support. Their analysis demonstrates that many Turkic, Iranian, and Caucasian languages fall into medium- or low-resource categories despite having millions of native speakers. Mikolov et al. (2013) showed that even relatively small corpora can produce useful word embeddings for NLP tasks, opening new possibilities for under-resourced language communities. More recently, large multilingual language models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) have demonstrated cross-lingual transfer capabilities, though their performance on Turkic languages remains below that of European languages due to the scarcity of pre-training data.

Studies of corpus development in comparable linguistic contexts — including Kazakh (Makhambetov et al., 2013), Kyrgyz (Washington et al., 2016), and Uyghur (Duval & Pitt, 2021) — highlight common bottlenecks: copyright restrictions on text collection, lack of standardised orthography, insufficient funding, and a shortage of trained corpus linguists. These studies consistently recommend international partnerships, open-source tooling, and government-backed language technology programmes as remedies.

Uzbek belongs to the southeastern branch of the Turkic family and exhibits agglutinative morphology, verb-final word order, and vowel harmony, though the latter has been substantially reduced in standard Uzbek under the influence of Tajik and Russian (Johanson, 1998; Sjoberg, 1963). The language has approximately 250 morphological suffixes, creating a rich inflectional and derivational system that poses significant challenges for automatic morphological analysis.

Soviet period (1924–1991) imposed a Cyrillic orthography that standardised some aspects of the written language but created a disconnect with earlier Arabic-script traditions. Independence in 1991 prompted a shift to a Latin-based alphabet, codified by law in 1993 and 1995, which has proceeded gradually alongside Cyrillic use in older publications (Landau & Kellner-Heinkele, 2001). This orthographic multiplicity means that any representative corpus of Uzbek must handle at least three scripts: Arabic (classical texts), Cyrillic (Soviet-era texts), and Latin (contemporary texts).

The Uzbek National Corpus (O'zbek Milliy Korpusi — OMK), initiated by a consortium of Uzbek universities in the 2010s, represents the most substantial attempt at large-scale Uzbek corpus construction. Published reports indicate a corpus of approximately 150 million tokens by 2022, with texts drawn from newspapers, fiction, academic writing, and legal documents (Matlatipov et al., 2022). However, public access remains restricted, annotation is limited to tokenisation and sentence segmentation, and documentation in English is sparse, limiting international collaboration.

Several smaller, domain-specific corpora have been developed for NLP research. Mirzayev et al. (2021) compiled a 12-million-token corpus of Uzbek news texts to train a transformer-based language model. Bakarov (2018) reviewed word-embedding resources for Uzbek and found that existing datasets were significantly smaller than those for Russian or Turkish. More recent work by Mansurov and Mansurov (2021) introduced a parallel Uzbek–English corpus of 1.5 million sentence pairs for machine translation research. Despite these efforts, no open, fully annotated reference corpus of Uzbek comparable in scope to the BNC or COCA currently exists.

MATERIALS AND METHODS

This study employs a convergent mixed-methods design (Creswell & Plano Clark, 2018), combining a systematic documentary review of Uzbek corpus resources with qualitative data from semi-structured interviews. The two strands were conducted concurrently and integrated at the interpretation stage to provide a comprehensive picture of corpus linguistics in Uzbekistan.

A systematic search of academic databases — including Scopus, Web of Science, Google Scholar, and the ACL Anthology (computational linguistics) — was conducted in January–March 2024. Search terms included: ("Uzbek" OR "Uzbekistan") AND ("corpus" OR "corpora" OR "natural language processing" OR "language resource"). Inclusion criteria were: (a) peer-reviewed publication or technical report; (b) primary focus on Uzbek language data; (c) publication between 2000 and 2024. After deduplication and screening of 312 initial records, 47 publications met all criteria and were included in the review.Each included publication was coded according to: corpus type (general, specialised, learner, parallel), corpus size (tokens), annotation level, availability (open access, restricted, unavailable), and primary research application. Coding was conducted independently by two researchers, with an inter-rater agreement of Cohen's $\kappa = 0.84$, indicating strong reliability. Ten participants — all active researchers or academic staff working in Uzbek corpus linguistics or closely related NLP fields at Uzbek universities or research institutes — were recruited through purposive sampling. Institutions represented included

Uzbekistan State World Languages University, the National University of Uzbekistan, Tashkent University of Information Technologies (TUIT), and the Institute of Language and Literature of the Academy of Sciences of Uzbekistan. Participant demographics: 7 male, 3 female; 5 with doctoral degrees, 5 with master's degrees; experience in the field ranging from 3 to 22 years. Interviews were conducted in Uzbek or Russian according to participant preference, audio-recorded with consent, and professionally transcribed. The interview protocol covered five thematic areas: (1) participants' current corpus-related projects; (2) available infrastructure and tools; (3) methodological challenges; (4) collaboration and funding; and (5) pedagogical use of corpus methods. Transcripts were analysed using thematic analysis (Braun & Clarke, 2006) with NVivo 14 software. Initial codes were generated inductively; these were subsequently organised into higher-order themes through iterative review.

Ethical approval was obtained from the research ethics committee of Uzbekistan State World Languages University (Protocol No. 7, 2024). All participants provided written informed consent and were assured of anonymity. Interview data are reported using participant codes (P1–P10) throughout.

RESULTS

The systematic review identified 18 distinct corpus resources or datasets relevant to Uzbek language research. Table 1 summarises the seven most significant resources by size. The Uzbek National Corpus (OMK) is the largest at an estimated 150 million tokens, but access requires institutional affiliation and formal request. Three corpora are fully open access: the Uzbek section of the Leipzig Corpora Collection (approximately 1 million tokens), the UzWiki corpus derived from Uzbek Wikipedia (approximately 60 million tokens), and the Uzbek Universal Dependencies treebank (UzUD, 6,500 sentences). The remaining resources are either restricted to their creating institutions or described only in academic publications without public release.Annotation levels across the reviewed corpora are predominantly basic. Only UzUD offers full syntactic annotation; five corpora include POS tags; and twelve are unannotated beyond tokenisation. No corpus provides

semantic annotation, discourse marking, or spoken-language transcription. Genre coverage is skewed towards news and official documents, with literary texts, spoken interaction, social media, and academic writing substantially underrepresented.

Analysis of the 47 included publications reveals four recurrent methodological challenges. First, orthographic heterogeneity is cited in 34 of 47 papers (72%) as a major complication, requiring script normalisation pipelines that may introduce errors or discard texts. Second, morphological complexity is addressed in 29 papers (62%): agglutinative Uzbek morphology creates data sparsity problems that affect both frequency-based and distributional analyses. Third, copyright and data availability are discussed in 22 papers (47%), with authors noting that much of the available digital text in Uzbek is locked behind paywalls or owned by state media. Fourth, tool availability is identified in 19 papers (40%): commercial and open-source NLP pipelines (e.g., spaCy, NLTK) lack mature Uzbek language models, obliging researchers to develop custom tools with limited resources. Thematic analysis of the ten interviews yielded four major themes: infrastructure deficit, funding fragmentation, limited corpus literacy, and emerging opportunities. All ten participants described the lack of a freely accessible, well-documented, richly annotated general corpus of Uzbek as the single greatest obstacle to advancing the field. P3 stated: 'We have the OMK but it is like a locked library. You can ask to enter, but the process is slow and many of my students give up.' P7 highlighted the absence of a spoken corpus: 'Everything we have is written. But language lives in speech. We cannot study colloquial Uzbek, regional variation, or prosody without a spoken resource.'Eight participants described the funding landscape as unpredictable and short-term. Most corpus projects are funded through one- to three-year grants from the Uzbekistan Agency for Innovative Development, with no guarantee of continuation. P1 observed: 'We built a 20-million word corpus of legal texts over two years. Then the grant ended. Now we have no money to annotate it or maintain the server.' International partnerships — primarily with Russian, Turkish, and US universities — were described by six participants as valuable but administratively burdensome. Seven participants noted that corpus methods are not systematically taught in Uzbek university linguistics programmes. P5 observed: 'Most of

my colleagues still do introspection-based linguistics. When I show them concordance lines or collocational profiles, they are surprised. It is not in our training.' This gap in corpus literacy limits demand for corpus resources and restricts the pool of researchers capable of building and using them. Despite these challenges, participants identified several positive developments. The inclusion of Uzbek in mBERT and XLM-R pre-training data, while small, provides a starting point for transfer learning. The growing Uzbek-language internet and social media presence — facilitated by the government's digitalisation agenda — offers new text sources. And a cohort of younger researchers trained abroad in computational linguistics (P2, P6, P10) is beginning to apply state-of-the-art methods to Uzbek data.

Only two of the ten participants (P4, P9) reported that corpus methods feature substantively in their undergraduate teaching. Three others (P1, P5, P8) include brief introductory sessions on corpus tools in graduate-level courses, while the remaining five reported no corpus component in their current teaching. Where corpus tools are used pedagogically, data-driven learning (DDL) approaches (Johns, 1991) are the most common application, particularly for grammar and collocation instruction. No participant reported using a purpose-built Uzbek pedagogical corpus, and most who use corpora in teaching rely on English-language resources (e.g., COCA or the British National Corpus) to demonstrate methodology.

DISCUSSION

The convergent findings from the documentary review and interviews paint a picture of a field with genuine scholarly momentum but significant structural constraints. The estimated 150-million-token size of the OMK, while superficially comparable to national corpora developed for languages of similar community sizes, masks critical deficits in annotation and accessibility. Comparison with the 100-million-token British National Corpus — which is fully POS-tagged, open access, and accompanied by extensive documentation (Burnard, 2007) — illustrates the gap. The finding that only one Uzbek corpus provides full syntactic annotation (UzUD, 6,500 sentences) is particularly significant given that syntactic treebanks are foundational for training parsers used in NLP

applications and advanced linguistic research. These findings resonate with comparative studies from Kazakh (Makhambetov et al., 2013) and Kyrgyz (Washington et al., 2016), suggesting that the infrastructure deficit in Uzbek corpus linguistics is not an isolated phenomenon but part of a broader regional pattern. The common thread across Central Asian Turkic languages is that corpus development has been initiated by academic institutions without sustained governmental investment in language technology infrastructure — a model that produces sporadic, incompatible, and often inaccessible resources.

The orthographic multiplicity identified in 72% of reviewed publications is particularly distinctive to the Uzbek context. Unlike most under-resourced languages, which face a single writing system, Uzbek corpus builders must navigate texts in Arabic, Cyrillic, and Latin scripts simultaneously. While automatic transliteration tools exist (e.g., the UzTranslit library), they introduce errors in cases of ambiguous grapheme-phoneme correspondences and do not resolve the deeper issue that Cyrillic-script Uzbek was itself a standardised variety that diverged in vocabulary and morphology from contemporary Latin-script Uzbek. A truly representative historical corpus of Uzbek would require dedicated annotation layers that flag orthographic variant rather than normalising across time periods.

The finding that corpus methods are systematically taught in only two of the ten participants' courses raises urgent questions about curriculum design in Uzbek university linguistics programmes. Research in corpus pedagogy consistently shows that data-driven learning approaches improve students' collocational competence, grammatical accuracy, and metalinguistic awareness (Boulton & Cobb, 2017; Flowerdew, 2015). The absence of such approaches from Uzbek linguistics education simultaneously limits the development of the next generation of corpus linguists and deprives learners of evidence-based language instruction.

Based on the converging documentary and interview evidence, the following recommendations are proposed for stakeholders at institutional, national, and international levels:

1. Open access mandate. The Uzbek National Corpus should be made fully accessible to researchers at Uzbek universities and, where copyright permits, to international scholars, following the model of the British National Corpus and the Sketch Engine corpus platform.

2. Systematic annotation. Funding should be prioritised for POS-tagging and dependency parsing of the existing OMK, leveraging the Universal Dependencies framework and adapting morphological analysers developed for related Turkic languages (e.g., Turkish TRmorph) to Uzbek.

3. Spoken corpus development. A dedicated spoken Uzbek corpus capturing regional, sociolectal, and genre variation is urgently needed. Methodological models from the spoken BNC2014 (Love et al., 2017) and the Russian National Corpus spoken sub-corpus provide relevant templates.

4. Curriculum integration. The national higher education authority should encourage the inclusion of corpus linguistics methods in undergraduate linguistics curricula, supported by Uzbek-language teaching materials and localised corpus tools.

5. International partnership. Engagement with initiatives such as the Endangered and Under-Resourced Languages (ELDP) program and the CLARIN European Research Infrastructure Consortium could provide technical expertise and funding leverage for Uzbek corpus development.

5.5 Limitations

This study has several limitations that qualify its conclusions. The interview sample of ten participants, while appropriate for qualitative thematic analysis, cannot capture the full diversity of perspectives within the Uzbek linguistics community. Purposive sampling towards researchers in Tashkent may underrepresent colleagues at regional universities. Furthermore, the documentary review was conducted in English and Russian, and may

have missed relevant publications in Uzbek that are not indexed in international databases. Future research should include larger-scale surveys and broader geographic coverage.

CONCLUSION

This article has presented the first comprehensive mixed-methods assessment of corpus linguistics in Uzbekistan, drawing on a systematic review of 47 publications and ten expert interviews. The findings demonstrate that while a foundation of corpus resources exists — most notably the Uzbek National Corpus and the Universal Dependencies treebank — the field is constrained by restricted access, limited annotation, orthographic heterogeneity, funding fragmentation, and insufficient corpus literacy among linguists and language teachers.

These challenges, though serious, are not insurmountable. Comparable under-resourced Turkic languages have made significant progress through a combination of international collaboration, open-source tool development, and government language technology investment. The growing digital infrastructure of Uzbekistan, the expansion of Uzbek-language internet content, and the emergence of a new generation of computationally trained linguists provide a favourable context for accelerated corpus development.

Corpus linguistics has the potential to transform Uzbek linguistic research, lexicography, language education, and NLP. Realising this potential requires coordinated action across universities, government agencies, and the international linguistics community. The recommendations advanced in this article provide a practical roadmap for that endeavour.

REFERENCES

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. Target, 7(2), 223–243.

Bakarov, A. (2018). A survey of word embeddings evaluation methods. arXiv preprint arXiv:1801.09536.

Biber, D. (1993). Representativeness in corpus design. Literary and Linguistic Computing, 8(4), 243–257.

Biber, D., Conrad, S., & Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge University Press.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. Language Learning, 67(2), 348–393.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101.

Burnard, L. (Ed.). (2007). Reference guide for the British National Corpus (XML edition). Oxford University Computing Services.

Carter, R., & McCarthy, M. (2017). Spoken grammar: Where are we and where are we going? Applied Linguistics, 38(1), 1–20.

Choudhury, M., & Jha, S. (2015). Assessment and interpretation of natural language processing tools for less-resourced languages. In Proceedings of LREC 2015 (pp. 4615–4619). ELRA.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of ACL 2020 (pp. 8440–8451). ACL.

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). SAGE Publications.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171–4186). ACL.

Duval, A., & Pitt, J. (2021). Building a corpus for Uyghur: Challenges and prospects. In Proceedings of the 5th Workshop on Language Technology for Language Documentation and Revitalization (LT4LangDoc 2021) (pp. 23–31). ACL.

Ethnologue. (2023). Uzbek. In Ethnologue: Languages of the world (27th ed.). SIL International. https://www.ethnologue.com/language/uzb/

Flowerdew, L. (2015). Learner corpus research and pedagogy. In S. Granger, G. Gilquin, & F. Meunier (Eds.), The Cambridge handbook of learner corpus research (pp. 413–434). Cambridge University Press.

Granger, S. (2003). The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research. TESOL Quarterly, 37(3), 538–546.

Johanson, L. (1998). The history of Uzbek. In L. Johanson & É. Á. Csató (Eds.), The Turkic languages (pp. 305–318). Routledge.

Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning. In T. Johns & P. King (Eds.), Classroom concordancing (ELR Journal, Vol. 4, pp. 1–16). University of Birmingham.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. Computational Linguistics, 29(3), 333–347.

Landau, J. M., & Kellner-Heinkele, B. (2001). Politics of language in the ex-Soviet Muslim states: Azerbayjan, Uzbekistan, Kazakhstan, Kyrgyzstan, Turkmenistan and Tajikistan. Hurst & Company.

Leech, G. (2005). Adding linguistic annotation. In M. Wynne (Ed.), Developing linguistic corpora: A guide to good practice (pp. 17–29). Oxbow Books.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics, 22(3), 319–344.

Makhambetov, O., Makazhanov, A., Yessenbayev, Z., Matlatipov, G., & Mamyrbayev, O. (2013). Assembling the Kazakh language corpus. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1022–1031). ACL.

Mansurov, B., & Mansurov, A. (2021). Uzbek-English parallel corpus. Zenodo. https://doi.org/10.5281/zenodo.4584558

Matlatipov, G., Mukhsimov, S., & Sobirov, A. (2022). Development of the Uzbek National Corpus: Current state and perspectives. Uzbek Linguistics and Literature, 4(1), 12–28. [In Uzbek]

McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.

McEnery, T., & Wilson, A. (2001). Corpus linguistics: An introduction (2nd ed.). Edinburgh University Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mirzayev, F., Yusupov, U., & Qodirov, B. (2021). UzBERT: A transformer-based language model for Uzbek. In Proceedings of the 2021 Conference on Asian Language Resources (ALR 2021) (pp. 58–65). ACL.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In Proceedings of LREC 2016 (pp. 1659–1666). ELRA.

Rissanen, M. (1994). The Helsinki Corpus of English texts: Classifying and coding the data. In M. Rissanen, M. Kytö, & M. Palander-Collin (Eds.), English in its social contexts (pp. 3–38). Mouton de Gruyter.

Sinclair, J. (1991). Corpus, concordance, collocation. Oxford University Press.

Sjoberg, A. F. (1963). Uzbek structural grammar. Indiana University Publications.

Washington, J. N., Salimzianov, I., Johnson, R., Strader, D., & Yeshkeyev, A. (2016). Initiating a Kyrgyz natural language processing initiative. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016) (pp. 3819–3823). ELRA.