European International Journal of Pedagogics (ISSN: 2751-OOOX)

Received: 30 April 2026 | Accepted: 26 May 2026 | Published: 30 June 2026

DOI: 10.55640/eijp-06-06-25

RESEARCH ARTICLE

Linguistic Foundations for The Automatic Identification of Prefixal Units in Uzbek

Narzullayeva Muborak Sherzod qizi
Gulistan State University, Gulistan, Uzbekistan

VOLUME: Vol.06 Issue06 2026

PAGE: 106-109

Abstract

The article sets out the linguistic foundations required for the automatic identification of prefixal units in Uzbek texts. Uzbek is an agglutinative language whose computational morphology has been built almost entirely around suffixation, and the preposed formatives anti-, mikro-, makro-, kiber-, eko-, no-, be- and ser- are therefore either segmented incorrectly or not segmented at all by existing analysers. The paper describes four sources of error: the homography of a prefix with the initial syllable of an unrelated root, the homography of a prefix with a free word, the graphic variation between solid, hyphenated and separated spelling, and the coexistence of the Latin and Cyrillic scripts with several competing transliteration conventions. On this basis a layered identification procedure is proposed, combining a closed lexicon of formatives with their combinatorial restrictions, a stop-list of homographic roots, a finite-state morphotactic component and a statistical confidence score derived from the productivity of the formative. The procedure is designed to be compatible with existing finite-state analysers of Turkic languages and with subword segmentation used in neural models.

KEYWORDS

Uzbek language, computational morphology, prefix identification, morphological segmentation, finite-state transducer, homography, subword units, corpus, natural language processing, lemmatisation.

INTRODUCTION

The automatic morphological analysis of Uzbek has developed as an analysis of suffixation. This orientation follows directly from the structural type of the language: Uzbek is agglutinative, its word forms are built by attaching an ordered sequence of suffixes to an invariable stem, and the overwhelming majority of grammatical and derivational information is therefore located to the right of the root [1]. Analysers built for Uzbek and for the neighbouring Turkic languages accordingly model the word as a stem followed by a chain of suffixal slots, and the description of morphotactics consists in specifying which slots may follow which. Within this architecture the left edge of the word is treated as fixed: whatever precedes the first suffix boundary is the stem, and no further analysis of it is attempted. As long as the vocabulary consists of inherited material this assumption is harmless, because inherited Turkic vocabulary contains no prefixes.

The assumption ceases to be harmless once the lexicon includes the prefixal formations that have entered Uzbek over the last century, and particularly over the last three decades. Units such as mikroqarz, kiberxavfsizlik, antikorrupsion, ekomahsulot, nostandart and serunum are morphologically complex on their left edge, and an analyser that treats the whole of kiberxavfsizlik as an unanalysable stem fails in three distinct ways. It cannot lemmatise the unit to a form that a dictionary contains, it cannot relate the unit to the base xavfsizlik with which it shares its principal meaning, and it cannot generalise from one member of a productive series to another, so that every new formation with kiber- must be

added to the lexicon by hand. The practical consequences appear in information retrieval, where a query for xavfsizlik fails to retrieve documents containing only kiberxavfsizlik, and in machine translation, where an unanalysed hybrid is treated as an unknown word.

The aim of the article is to establish the linguistic foundations on which an automatic identification of prefixal units in Uzbek can be built. The tasks are to define the inventory of formatives to be recognised, to describe the sources of ambiguity that make naive left-edge matching unreliable, to formulate the combinatorial restrictions that reduce this ambiguity, and to propose an identification procedure compatible with the architectures currently used for Turkic morphology. The material consists of prefixal formations attested in Uzbek press and technical texts; the methods are morphemic analysis, distributional analysis of the formatives and a formal description of the identification rules in terms suitable for implementation.

METHOD

The first task of any identification procedure is to fix the inventory of units to be recognised, and for prefixal formatives in Uzbek this inventory is closed and comparatively small, which is a considerable advantage. It comprises three groups. The Iranian group contains be-, no-, ba-, bar-, ser-, kam- and ham-, all of them long assimilated and all of them recorded in the morpheme dictionaries of the language [2]. The international group contains anti-, arxi-, avto-, bio-, geo-, giper-, gipo-, dez-, infra-, kvazi-, makro-, mikro-, mono-, multi-, neo-, poli-, post-, psevdo-, re-, sub-, super-, tele-, trans-, ultra- and ekstra-. The recent group contains kiber-, veb-, eko-, nano-, smart-, onlayn- and the abbreviated e-. Because the inventory is closed, it can be stored as a finite lexicon rather than being learned, and the identification problem reduces from segmentation in the general sense to the disambiguation of a small set of known strings at the left edge of the word.

The principal source of error is the homography of a prefixal formative with the initial syllable of an unrelated root. Every formative in the inventory consists of a short and phonotactically ordinary sequence, and Uzbek contains numerous roots that begin with exactly that sequence without containing the formative. The formative no- is homographic with the initial syllable of nok, nom, nozik, nogiron and noyabr; ba- with that of bahor, balandlik, bahs and bank; ser- with that of servis, serial and sertifikat; kam- with that of kamon, kamar and kamera; ham- with that of hamma, hammom and hamyon; re- with that of reja, rejim and reklama; mono- with that of monitor; post- with that of pochta in some transliterations. A procedure that segments on string matching alone will therefore analyse nogiron as no- plus a non-existent base and servis as ser- plus vis, producing errors on precisely those high-frequency words where errors are most damaging.

The homography problem is not solvable by string manipulation and requires lexical knowledge, but it is solvable cheaply because the errors are systematic. The decisive test is the existence of the residue as an independent lexical item: a segmentation is admissible only if what remains after the removal of the formative is itself a word of Uzbek recorded in the lexicon. On this test nogiron is rejected, because giron is not a word; nostandart is accepted, because standart is; servis is rejected, because vis is not a word; serunum is accepted, because unum is. The test requires a reasonably complete lexicon of stems, which is available for Uzbek, and it eliminates the great majority of false positives at negligible computational cost. A residual set of genuinely ambiguous cases remains, in which the residue is a word but the resulting meaning is not compositional, and these are best handled by an explicit stop-list of frequent exceptions rather than by additional rules.

A second source of error is the homography of a formative with a free word of Uzbek, which produces ambiguity of a different kind. The formatives kam- and ham- correspond to the free words kam meaning few and ham meaning also, and the string kam at the left edge of a longer sequence may therefore represent either the formative or a separate word that has been written without a space. The same difficulty arises with bir-, o‘z- and tez- in occasional formations. Here the residue test is insufficient, because both analyses may leave a genuine word, and the resolution depends on the semantic relation between the two elements and on the orthography of the source text. A practical solution is to treat such cases as ambiguous and to pass both analyses to the next stage, allowing a syntactic or statistical component to choose between them, rather than forcing a decision at the morphological level.

The third source of error is graphic variation. The same prefixal unit is attested in Uzbek in three spellings, solid, hyphenated and separated, and the choice among them is not fixed by norm for the recent formatives. A corpus of contemporary press texts will contain ekoturizm, eko-turizm and eko turizm, and an analyser that recognises only the solid form will treat the other two as a hyphenated compound and as two separate tokens respectively. Normalisation must therefore precede identification: the tokeniser should be instructed that a string from the formative lexicon followed by a hyphen, or followed by a space and a word from the stem lexicon, is a candidate single token. This normalisation must be reversible, so that the original form can be restored in the output, and it must be restricted to the closed formative inventory, since a general rule joining hyphenated sequences would destroy genuine compounds and reduplications such as ora-chora and yaxshi-yomon.

The fourth source of error is script and transliteration variation. Uzbek is written in both the Latin and the Cyrillic alphabets, and within the Latin alphabet the characters representing the vowels of o‘ and g‘ are rendered in practice

by at least three different sequences, the typographic apostrophe, the straight apostrophe and the modifier letter. A formative such as o‘z- and stems containing these characters therefore appear in several graphic shapes that are distinct as strings but identical as words. For prefix identification specifically, the relevant consequence is that the residue test will fail if the lexicon stores one variant and the text contains another. The remedy is a normalisation layer applied to both the lexicon and the input, which maps all variants of each character to a single internal representation before any matching is attempted; this layer is required in any case for other components of Uzbek text processing and imposes no additional cost [3].

Beyond disambiguation, the identification procedure requires a statement of the combinatorial restrictions of each formative, since these restrictions both improve precision and make the analysis linguistically informative. Three types of restriction are relevant. Categorial restrictions specify the part of speech of admissible bases: anti- and eko- attach to nouns and to adjectives derived from them, whereas re- attaches predominantly to verbal nouns of international origin. Etymological restrictions specify the origin of admissible bases: psevdo-, kvazi- and arxi- combine almost exclusively with borrowed bases, whereas mikro-, eko- and kiber- combine freely with native and assimilated ones. Semantic restrictions specify the conceptual domain: giper- and gipo- are confined to medical and physical terminology, and their appearance before a base outside these domains is more likely to indicate a false positive than a genuine neologism. These restrictions can be encoded as feature constraints in a finite-state lexicon of the kind used for Turkic morphology [4].

The architecture that follows from these observations is layered. The first layer normalises the script and the graphic variants of the input. The second layer performs candidate generation by matching the left edge of the token against the closed formative lexicon, producing all possible analyses. The third layer applies the residue test against the stem lexicon and the stop-list of homographic roots, discarding inadmissible candidates. The fourth layer checks the combinatorial restrictions of the surviving candidates and assigns each a confidence score derived from the productivity of the formative and the frequency of the base. The fifth layer passes the analysed stem, now correctly identified as prefix plus base, to the existing suffixal analyser, which proceeds as before. This design has the practical advantage that it requires no modification of the suffixal component: prefix identification is inserted as a preprocessing stage, and existing finite-state descriptions of Uzbek morphology remain valid [5].

The relation between this procedure and the subword segmentation used in current neural models deserves separate comment, because the two are often assumed to make each other unnecessary. Statistical subword methods such as byte-pair encoding segment words into frequent character sequences without linguistic knowledge, and they do in fact isolate frequent formatives such as mikro and anti as separate units simply because these sequences are frequent [6]. The segmentation they produce, however, is not morphological: it is determined by frequency rather than by meaning, it varies with the training corpus, and it splits infrequent bases into arbitrary fragments. For tasks that require the identification of the base as a lexical item, such as lemmatisation, dictionary lookup, terminological extraction and information retrieval, statistical segmentation is therefore not a substitute for morphological analysis, although it is a useful complement to it. Unsupervised morphology learning occupies an intermediate position and can be used to propose candidate formatives for inclusion in the lexicon, which is valuable for a language whose prefixal inventory is still expanding [7].

Evaluation of a prefix identification component requires a resource that does not yet exist for Uzbek, namely a manually annotated set of prefixal formations covering all three layers of the inventory together with a matched set of homographic negatives such as nogiron, servis and kamera. Without the negatives, accuracy figures are uninformative, because the difficulty of the task lies entirely in the rejection of false positives; a procedure that accepts every left-edge match will score highly on a positive-only test set and will be unusable in practice. The construction of such a resource is a realistic task, since the formative inventory is closed and the homographic roots can be extracted mechanically from an existing dictionary by matching each formative against the initial substring of every entry. This is the most immediate practical step towards the automatic treatment of prefixation in Uzbek [8].

The identification of prefixal units is not an end in itself, and the design of the component should be governed by the tasks it serves. For lemmatisation and dictionary lookup the requirement is that the base be returned as an independent lexical item, so that a query for xavfsizlik retrieves documents containing kiberxavfsizlik; here recall matters more than the correct labelling of the formative. For terminological extraction the requirement is the opposite: the formative must be identified and labelled, because it is the formative that assigns the candidate term to a conceptual domain, and a system that recognises giper- can classify an unknown medical term without having seen it before. For machine translation the requirement is different again, since the formative must be mapped to its counterpart in the target language, and the mapping is not always one to one: Uzbek no- corresponds to English un-, in-, non- and dis- according to the base, and the choice can only be made by consulting the target lexicon [9]. A single identification component can serve all three tasks provided that it outputs the full analysis, formative and base and their labels, rather than a segmentation alone.

A remark should be added about the diachronic dimension, which is unusual in computational morphology but is unavoidable for Uzbek. The inventory of prefixal formatives is closed at any given moment but is not fixed over time: kiber-, veb- and nano- were absent from Uzbek a generation ago, and further formatives will enter as new domains of vocabulary are borrowed. A lexicon compiled today will therefore be incomplete in a decade, and the component must be designed so that adding a formative requires only an entry in the lexicon and its combinatorial restrictions, not a modification of the rules. This argues for a clean separation between the declarative lexicon and the procedural layers, which is in any case the design principle of finite-state morphological description. It also argues for periodic corpus-driven review, in which candidate formatives are proposed by unsupervised segmentation and verified against the criteria of boundness, recurrence and semantic stability that define a prefix in the descriptive grammar of the language [10].

CONCLUSION

Automatic morphological analysis of Uzbek has been built on the assumption that the language is exclusively suffixing, and this assumption no longer matches the vocabulary of the modern language. Prefixal formations built with the Iranian formatives be-, no- and ser-, with the international formatives anti-, mikro- and post-, and with the recent formatives kiber-, eko- and veb- are numerous, productive and growing, and an analyser that treats their left edge as an unanalysable stem loses the lemma, the derivational relation and the ability to generalise across a productive series.

The task is nevertheless tractable, because the inventory of formatives is closed and small. Four sources of error have been identified and each admits a specific remedy: homography with the initial syllable of an unrelated root is resolved by the residue test against a stem lexicon supplemented by a stop-list; homography with a free word is resolved by passing both analyses to a later stage rather than deciding prematurely; graphic variation between solid, hyphenated and separated spelling is resolved by a reversible normalisation restricted to the formative inventory; script and transliteration variation is resolved by a normalisation layer applied to lexicon and input alike. Combinatorial restrictions of a categorial, etymological and semantic kind further increase precision and can be encoded as feature constraints in a finite-state lexicon.

The proposed procedure is layered and is designed to be inserted before an existing suffixal analyser rather than to replace it, so that current descriptions of Uzbek morphology remain usable without modification. Statistical subword segmentation is a complement to this procedure and not a substitute for it, since frequency-based segmentation does not identify the base as a lexical item. The immediate practical prerequisite is an annotated evaluation set containing both prefixal formations and homographic negatives, without which the accuracy of any identification method cannot be assessed. Further work should extend the formative lexicon as new elements enter the language and should test the procedure on the two scripts in which Uzbek is written.

REFERENCES

1. Hojiyev A. O‘zbek tili so‘z yasalishi. – Toshkent: O‘qituvchi, 1989. – 168 b.
2. G‘ulomov A., Tixonov A., Qo‘ng‘urov R. O‘zbek tili morfem lug‘ati. – Toshkent: O‘qituvchi, 1977. – 464 b.
3. Jurafsky D., Martin J. H. Speech and Language Processing. 3rd ed. – Stanford: Stanford University, 2023.
4. Oflazer K. Two-level Description of Turkish Morphology. // Literary and Linguistic Computing. – 1994. – Vol. 9, No. 2. – P. 137–148.
5. Beesley K. R., Karttunen L. Finite State Morphology. – Stanford: CSLI Publications, 2003. – 611 p.
6. Sennrich R., Haddow B., Birch A. Neural Machine Translation of Rare Words with Subword Units. // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. – Berlin, 2016. – P. 1715–1725.
7. Creutz M., Lagus K. Unsupervised Models for Morpheme Segmentation and Morphology Learning. // ACM Transactions on Speech and Language Processing. – 2007. – Vol. 4, No. 1. – P. 1–34.
8. Koskenniemi K. Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. – Helsinki: University of Helsinki, 1983. – 160 p.
9. Booij G. The Grammar of Words: An Introduction to Linguistic Morphology. 2nd ed. – Oxford: Oxford University Press, 2007. – 353 p.
10. Rahmatullayev Sh. Hozirgi adabiy o‘zbek tili. – Toshkent: Universitet, 2006. – 464 b.