MODELING MORPHOLOGICAL ANALYSIS BASED ON WORD-ENDING FOR UZBEK LANGUAGE

Ulugbek Salaev
PhD student at Urgench State University

https://doi.org/10.5281/zenodo.10155225

Abstract. *Uzbek, an agglutinative language, forms words by combining affixes with roots, utilizing inflectional endings for various morphological features. This property makes a large number of combinations of word ending, and greatly increases the word-vocabulary size, and data sparseness problems for statistical models. This paper discusses a morphological analyzing model which includes stemming, lemmatizing and extraction of morphological information considering morpho-phonetic exceptions. A main point of the model involves developing a complete set of word-ending with assign morphological information, and additional datasets for morphological analysis. The proposed model was evaluated using a curated test set comprising 5.3K words. It achieved a word-level accuracy over 91%, as determined through manual verification of stem, lemma, and morphological feature corrections conducted by linguistic experts. The created tool based on the proposed methodology is available as an open-source Python package, as well as a web-based application including a public API.*

Keywords: *uzbek language, morphological analyzing, morphological segmentation, stemming, lemmatizing, bound morpheme, inflectional ending.*

Introduction. Computational linguistics integrates human natural language modeling through rule-based approaches with statistical, Machine Learning, and Deep Learning models.

The fact that some languages extensively use suffixes and prefixes to convey grammatical meaning (e.g. subject-verb agreement) poses a challenge to most current human language technology. Suffixes and prefixes in such languages can more generally be called morphemes, which are defined as the meaningful subparts of words. The rules that languages use to combine morphemes, together with the actual morphemes that they use, are both referred to as a language's morphology. Languages which make extensive use of morphemes to build words are said to be morphologically-rich. The Uzbek language like other Turkic languages has rich morphology and many exception cases in morphological structure of the words.

Morphology focuses on how the parts of a word, like stems, prefixes, and suffixes, are organized or changed to convey different meanings. There are two main categories of morphology: inflectional morphology and derivational morphology. These two categories each hold their own importance in various aspects of NLP. Inflectional morphemes, typically suffixes, are appended to the end of a word to convey grammatical nuances. These inflectional endings serve to modify the word’s meaning while retaining its fundamental core meaning. In this project we explore morphologic analyzing model based on the complete set of word ending that formed with only inflectional morphemes. The morphological analysis model includes stemming, lemmatizing and morphological segmentation methods along with morphological information extraction.

Morphological analysis in Uzbek involves the figure out of complicated word structures, encompassing prefixes, suffixes, and root forms. The highly inflectional nature of the language necessitates a nuanced approach to feature extraction and understanding contextual semantics.

Neural network models, with their capacity to capture complex patterns and contextual properties, hold the potential to better in this demanding linguistic landscape.

Literature review. The work [1] has been addressed by developing morphological analyzers that segment words into sequences of morphemes or syllables in Uygur language. The morpheme segmentation process is central to the creation of a comprehensive Uyghur language corpus. While a supervised approach combined with rules and statistical learning algorithms, effectively handles morpheme segmentation, the value of linguistic morphemes in enhancing word coverage, reducing lexicon size. Furthermore, these morphological analyzers are designed to handle both standard and surface forms, accommodating phonetic alterations and complex morphological changes.

Storing all surface words in a dictionary for a morphologically rich Turkic language has long been a challenge and imposing constraints on the widespread application of neural machine translation (NMT). The paper [2], [3] addresses this limitation by introducing an innovative approach to morphological segmentation for Turkic languages, rooted in the concept of the complete set of endings (CSE). The CSE-based segmentation method, as demonstrated for Kazakh, Kyrgyz, and Uzbek languages, effectively diminishes the vocabulary size within source corpora. The computational NMT experiments, notably focusing on the Kazakh language, reveal compelling results. When compared to byte-pair encoding (BPE)-based segmentation, the CSE-based approach enhances the bilingual evaluation understudy score for Kazakh–English and English–Kazakh pairs. In the context of Uzbek language morphological analysis, several models using finite state machines (FSM) have been proposed. In [4], the authors employ predefined grammatical rules to extract morphological information from given words using automata. However, it’s important to note that this methodology utilizes a lexicon of word forms as a fundamental database, in contrast to the utilization of grammatical, morphotactic, and morphology rules. In [5], which is one of the works on the morphological analysis of Uzbek language words based on the finite state machine (FSM) model, explores a stemming task with the extraction relevant morphological information. Meanwhile, for the Uzbek language, the morphotactic rules [6] and morphological analysis models [7], [8] have been developed. The Uzbek words are considered to have a rather complex structure, and there are many exceptional cases. Uzbek language presents a challenge due to the presence of homonymy and synonymy of affixes, vowel harmony which can introduce errors in morphological segmentation. This also may cause the decreasing accuracy of morphological analysis models based on FSM.

Methodology. In this section, we present the detailed description of the Uzbek morphological analyzing model. The model utilized completes set of word-ending dataset, and supplementary datasets and include rule-based stemming, lemmatization and morphological analyzing methods. The proposed methodology overview presents in Figure 1.

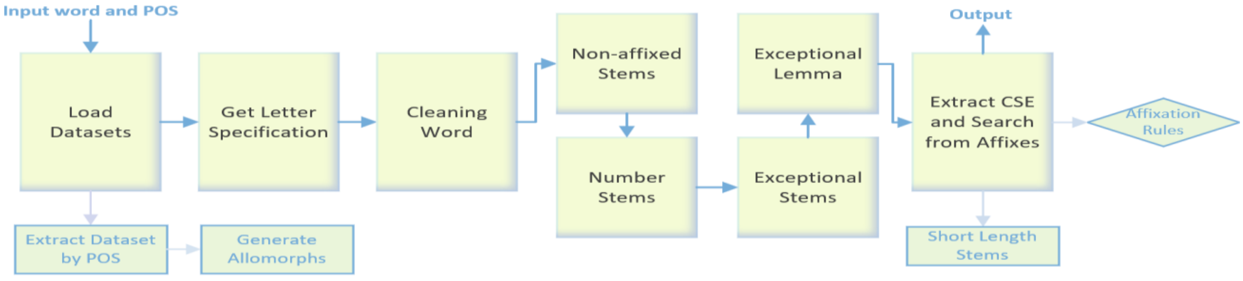


Figure 1. An overview of the proposed model

The initial step of the proposed methodology involves the developing complete set of ending (CSE) dataset of the language. Morphemic part of the word will be divided into two parts, derivational and inflectional morphology. Inflectional morphology encompasses sequences of lexical and syntactical suffixes. Derivational morphemes change the part of speech of a word. In this work we maintain inflectional part of the words. During the forms of the dataset, we collected all possible word endings; morphological information associated with each word ending is provided. This morphological information includes various linguistic attributes such as part of speech (POS), tense, possession, copula, singular or plural form, question form, case, and more. The dataset of word endings is generated by linguistic experts to ensure the accuracy and correctness of the morphological information. By the analyzing of morphological structure of the words, database structure developed for store all necessary information for the model of morphological analyzing. The database structure and relationships of the tables presents in Figure 2.

The reason of vowel harmony and affixation rules, a suffix in Uzbek can have multiple allomorphs. So, in the dataset of word endings, marked specific denotations to capture variations and allomorphs. In order to describe allomorphic suffixes, we use the following notation from the previous work [5]:

G:{g,k,q}; K:{g,k}; Q:{g,g‘,k,q}; Y:{a,y}; T:{t,d}; (): the letter enclosed within parentheses can be omitted.

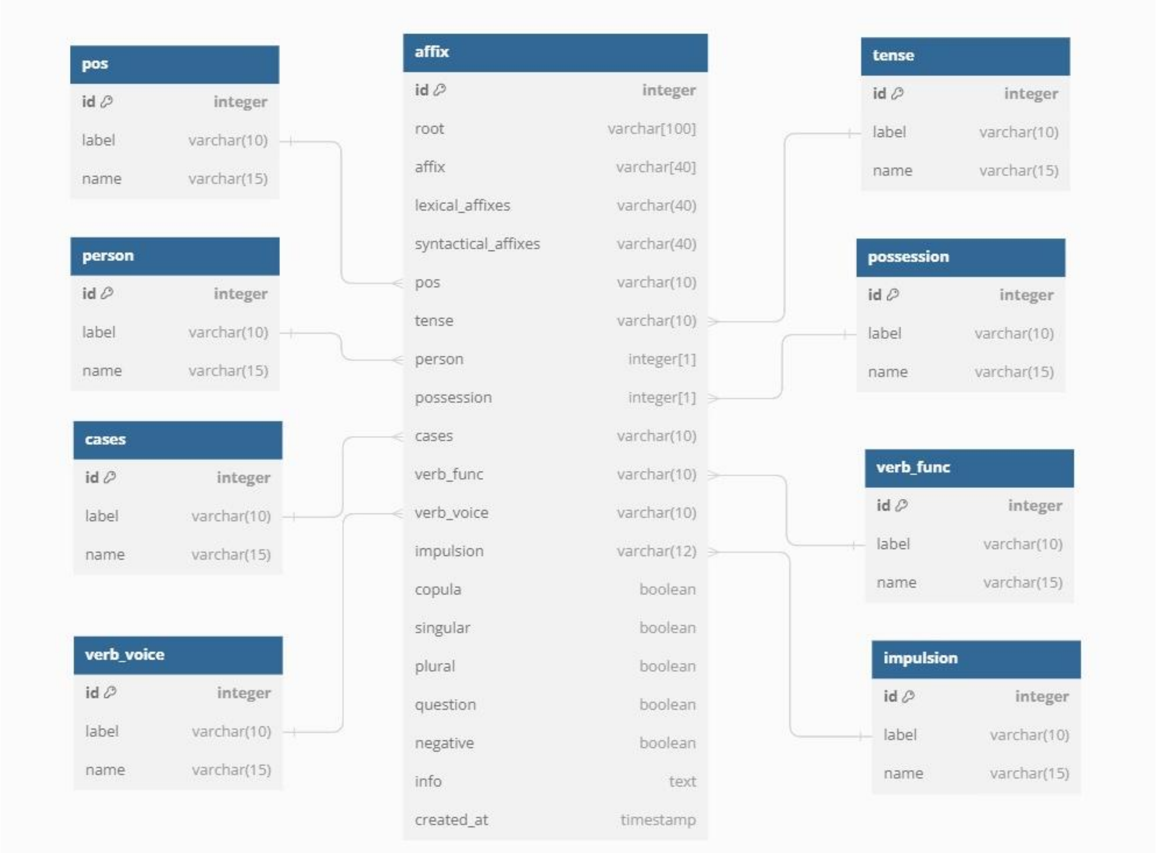
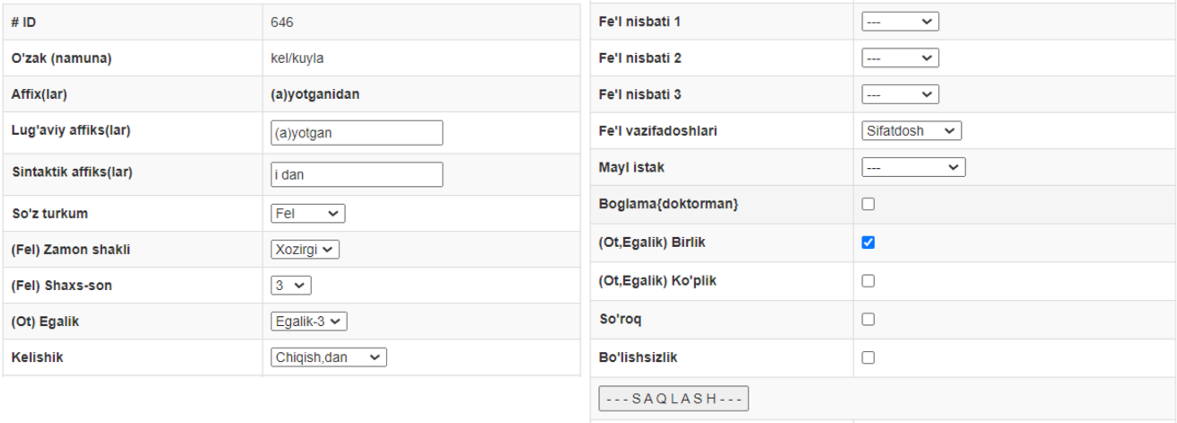


Figure 2. Structure and relationship of the CSE dataset.

For the input process, we developed a web-based application where each linguist expert is given a unique username and password, where they can access the website and input new word-ending with the relevant morphological information. General user interface of the forms database can be seen in Figure 3.



# ID	646
O'zak (namuna)	kel/kuyla
Affix(lar)	(a)yotganidan
Lug'aviy affiks(lar)	(a)yotgan
Sintaktik affiks(lar)	i dan
So'z turkum	Fel
(Fel) Zamon shakli	Xozirgi
(Fel) Shaxs-son	3
(Ot) Egalik	Egalik-3
Kelishik	Chiqish,dan

Fe'l nisbati 1	---
Fe'l nisbati 2	---
Fe'l nisbati 3	---
Fe'l vazifadoshlari	Sifatdosh
Mayl istak	---
Boglama{doktorman}	☐
(Ot,Egalik) Birlik	☑
(Ot,Egalik) Ko'plik	☐
So'roq	☐
Bo'lishsizlik	☐

Figure 3. Web Tool Interface for CSE Dataset Entry

A significant portion of inflectional word endings correctly corresponds to verb and noun as grammatical function of a word. In the dataset, 1205 and 150 entries are verbs and nouns, respectively. Additionally, the dataset comprises 22 entries of numerals, 10 of adjectives, 20 of pronouns, and 10 entries of adverbs.

Additionally, we formed five datasets to increase accuracy and performance of the methodology.

Exceptional stems (*dollar, tashqari, …*), List of the stems which exist some affix tail in it.

Non-affixed stems (*va, lekin, yana, …*), list of the stems which never append any affix to it.

Number stems (*bir, ikki, ellik, …*).

Short length stems (length<=2) (*u, bu, ot, un, …*).

Exceptional Lemma (bitta,bir,ta / singli,singil,i).

In the next step of the methodology, allomorphs will be generated from the affix column in CSE dataset for each item. Subsequently, the modeling process will incorporate alphabet specifications that are used to check morphological rules. Following this, the word inputs will transfer to a cleaning process, which includes converting all characters to lowercase and replacing specific special characters as required. This ensures that the input words are appropriately prepared for subsequent processing and rule checking.

Before determining a word-ending of the inputted word, we check predefined stems in additional created datasets, which ensures that the model runs fast and has a low error rate. The word ending is extracted, providing an additional layer of morphemic information. To enhance the accuracy of morphological analysis, affixation rules are applied during the extraction of word endings. The extracted word ending is then compared against the CSE dataset to identify matches. Upon finding a match, relevant morphological information is assigned to the result set. When the predicted stem length is less than two characters, an additional verification step is implemented. The short length stem dataset is consulted to confirm the correctness of the stem. This validation process ensures the accuracy of short stems, contributing to the overall reliability of the morphological analysis. These processing steps results returns a list included stem, lemma and morphological information for the given word and POS tag.

Results and Discussions. As evaluation purpose, the proposed model has been analyzed using the constructed corpus data for morphological segmentation method. The constructed corpus for testing was utilized, consisting of 40 documents sourced from a news platform (daryo.uz), categorized into four groups with ten documents each. This corpus encompasses a total of 11,952 words, with 5,288 unique words, accounting for 44.24% of the total words. The words were annotated by linguist expert with correct segmentation. In the evaluation process, the performance time of the model take about 1000 words for a second. The model's output comparing for each word with the actual expected output. The model's outputs were categorized into five distinct cases, one case of them indicating correct predictions, while the other four indicating various types of errors encountered during the evaluation. In Table 1, the evaluation results of the model on the compiled corpus demonstrate the categorization of outputs into the cases. In the table, the surface word is defined as the sample *kitobimdan* (*of my book*), it is divided into three morphemes: *kitob* [stem], -(*i*)*m* [first person possession], and -*dan* [ablative case].

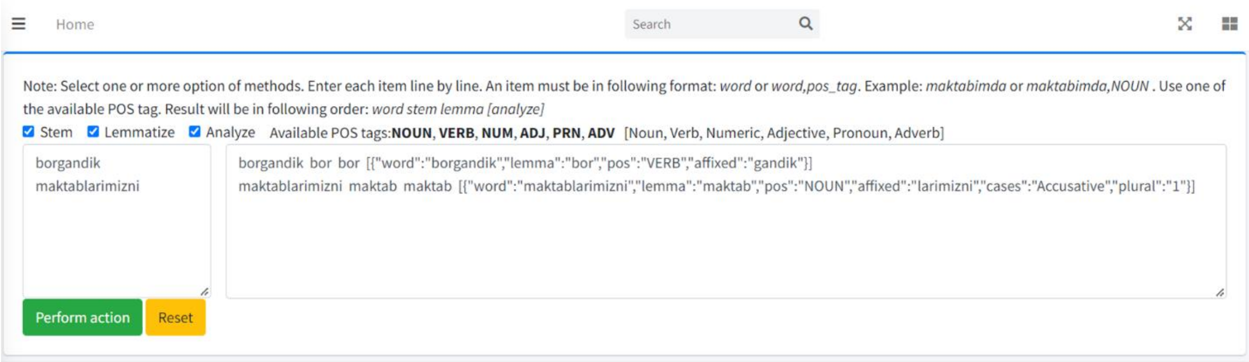
Table 1. Evaluation Results of Morphological Segmentation Method

Cases	Sample	Token	%
1. Correct prediction	*kitobimni > kitob*	4821	91.2%
2. No affixes were striped while affix(es) exists	*kitobimni > kitobimni*	24	0.5%
3. Affix(es) were striped from the stem, while stem has affix(es)	*kitobimni > kito*	131	2.5%
4. Partially striped affixes	*kitobimni > kitobim*	158	3.0%
5. Striped affix(es) from the stem, while has no affix(es)	*kitob > kito*	154	2.9%
		5288	

Most of the errors occur due to the existing of short-length word-endings in stems (3rd, 5th cases) and all options are not yet covered in the CSE dataset (4th case).

The Python tool created for this work is openly-accessible, and also can be easily installed, using the following command that is popular for the Python community: pip install UzMorphAnalyser

To demonstrate the model performance a web interface[1] was created, the web tool can be seen in Figure 4. There is also a public API system[2] for integrating the model into other software, and more detailed information about it can be found at the project's GitHub repository[3].



Note: Select one or more option of methods. Enter each item line by line. An item must be in following format: *word* or *word,pos_tag*. Example: *maktabimda* or *maktabimda,NOUN* . Use one of the available POS tag. Result will be in following order: *word stem lemma [analyze]*

☑ Stem ☑ Lemmatize ☑ Analyze Available POS tags:**NOUN**, **VERB**, **NUM**, **ADJ**, **PRN**, **ADV** [Noun, Verb, Numeric, Adjective, Pronoun, Adverb]

```
borgandik
maktablarimizni
```

```
borgandik bor bor [{"word":"borgandik","lemma":"bor","pos":"VERB","affixed":"gandik"}]
maktablarimizni maktab maktab [{"word":"maktablarimizni","lemma":"maktab","pos":"NOUN","affixed":"larimizni","cases":"Accusative","plural":"1"}]
```

Figure 4. Web interface of Morphological Analyzer Model Performance

Conclusion. This study conducts a comprehensive morphological analysis of the Uzbek language, encompassing various tasks related to word structure and morphological properties. The methodology entails the development of an inflectional ending dataset, annotated with

[1] https://nlp.urdu.uz/?menu=analyzer
[2] https://api.urdu.uz/docs
[3] https://github.com/UlugbekSalaev/UzMorphAnalyser

morphological information by linguistic experts. Most bound morphemes, particularly inflectional endings, are generated through nouns and verbs, which play a role in modifying the grammatical properties of words. The evaluation process involves stemming, lemmatization, and morphological analysis, resulting in an impressive accuracy rate over 91% for stem and lemma validation. Additionally, we presented a Python code, a web tool, and an API created for using proposed model.

Our future work aims to enhance the output quality of the current tool by expanding its coverage of inflectional endings, incorporating additional morphotactic rules, and integrating a pretrained neural language model. Additionally, we plan to develop a comprehensive pipeline capable of performing essential NLP tasks for the Uzbek language, including morphological generation, POS tagging, and syntactic parsing in a foreseen future.

REFERENCES

1. M. Ablimit, T. Kawahara, A. Pattar, and A. Hamdulla, “Stem-Affix based Uyghur Morphological Analyzer,” *International Journal of Future Generation Communication and Networking*, vol. 9, no. 2, 2016, doi: 10.14257/ijfgcn.2016.9.2.07.
2. U. Tukeyev, A. Karibayeva, and Z. h. Zhumanov, “Morphological segmentation method for Turkic language neural machine translation,” *Cogent Eng*, vol. 7, no. 1, p. 1856500, 2020, doi: 10.1080/23311916.2020.1856500.
3. A. and T. A. and A. D. Tukeyev Ualsher and Karibayeva, “Universal Programs for Stemming, Segmentation, Morphological Analysis of Turkic Words,” in *Computational Collective Intelligence*, L. and M. I. and T. B. Nguyen Ngoc Thanh and Iliadis, Ed., Cham: Springer International Publishing, 2021, pp. 643–654.
4. I. I. Bakaev and R. I. Bakaeva, “Creation of a morphological analyzer based on finite-state techniques for the Uzbek language,” in *Journal of Physics: Conference Series*, 2021. doi: 10.1088/1742-6596/1791/1/012068.
5. M. Sharipov and U. Salaev, “Uzbek affix finite state machine for stemming,” *arXiv preprint arXiv:2205.10078*, 2022.
6. Khamroeva Shahlo, “MORPHOTACTIC RULES IN THE MORPHOLOGICAL ANALYZER OF THE UZBEK LANGUAGE,” *Middle European Scientific Bulletin*, vol. 6, 2020, doi: 10.47494/mesb.2020.6.112.
7. N. Abdurakhmonova, I. Alisher, and R. Sayfulleyeva, “MorphUz: Morphological Analyzer for the Uzbek Language,” in *Proceedings - 7th International Conference on Computer Science and Engineering, UBMK 2022*, 2022. doi: 10.1109/UBMK55850.2022.9919579.
8. G. Matlatipov and Z. Vetulani, “Representation of Uzbek morphology in prolog,” in *Aspects of Natural Language Processing*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 83–110. doi: 10.1007/978-3-642-04735-0_4.