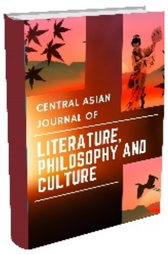


CENTRAL ASIAN JOURNAL OF LITERATURE, PHILOSOPHY AND CULTURE

eISSN: 2660-6828 | Volume: 04 Issue: 12 Dec 2023
https://cajlpc.centralasianstudies.org

Methods of Lemmatizing Various Structured Words in Uzbek Language

Zilola Xusainova
Doctoral candidate at TSUULL

Received 4th Oct 2023, Accepted 5th Nov 2023, Online 14th Dec 2023

ABSTRACT

Lemmatization is the process of finding the main morphological form (lemma) of a word. This is an important first step in solving many Natural Language Processing (NLP) and Information Retrieval (IR) problems. Lemmatization is a complex task due to the lexical meaning of the Uzbek language due to the richness of its morphology and agglutinative aspects. This article presents methods of lemming words with different structures in the Uzbek language.

KEYWORDS: Lemmatization, stemming, search engines, morphological analysis, words with different structures, compound word, double word, repeated word.

Introduction

To enhance the efficiency of information retrieval from the Uzbek language corpus and obtain precise results, it is necessary to shorten the corpus with initial focus on reducing its size significantly [B. Elov, Sh. Hamroyeva, D. Elova. 2022]. The first phase of the information retrieval process involves reducing the corpus size [B.Elov, Sh.Hamroyeva, D.Elova. 2022]. Initial tasks such as removing stop words, stemming, and lemmatization are carried out to improve the accuracy of information retrieval. In stemming, the inflected forms of words are produced by removing prefixes or suffixes [B.B.Elov, Sh.M.Hamroyeva, O.X.Abdullayeva, Z.Y.Xusainova, N.U.Xudayberganov. 2023:43]. Lemmatization is the process of identifying the canonical forms of words, providing a normalized dictionary form with a single or related meaning. It plays a crucial role in various Natural Language Processing (NLP) systems [I.Boban, A.Doko, S.Gotovac. 2020:349]. Lemmatization - serves as a fundamental stage for many applications, contributing to the essential functioning of NLP systems and facilitating a deeper understanding of natural language. Lemmatization, unlike stemming provides precise and necessary results [I.Boban, A.Doko, S.Gotovac. 2020:349].

Lemmatization is the process of grouping different inflectional forms of a word for analysis as a single element.It is a crucial step for many applications that require a deeper understanding of natural language, offering significant improvements over stemming [I.Boban, A.Doko, S.Gotovac. 2020:349]. It is used to identify the dictionary form or lemma of a word, representing its normalized shape. A lexeme encompasses the combination of all forms having the same meaning (referred to as the headwords of a dictionary), while a lemma is selected as the pivotal form that expresses the lexeme. Both stemming and lemmatization processes are considered crucial stages in obtaining information for tasks related to language understanding.This process

Copyright (c) 2023 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY).To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

is vital for finding key words in search systems and reducing the size of index files. Creating an accurate and complete natural language dictionary is essential for developing a lemmatizer.

In this process, dictionaries, morphological analysis, and word categorization are commonly used. The key difference between stemming and lemmatization lies in the fact that lemmatization considers context and transforms a word into its lemma, while stemming often cuts off final (sometimes initial) characters, leading to inaccuracies in meaning and spelling. In many cases, this leads to incorrect meaning and spelling errors. In stemming, a heuristic method is typically used to remove affixes from words.

Table 1. *The process of stemming and lemmatization*

Wordform	Stem	Lemma	Root
Kelajagimiz	Kelajag	Kelajak	Kelajak
o‘quvchilar	o‘quv	o‘quvchi	o‘qi
borib ketdi	bor ket (2ta)	borib ketmoq	bor ket (2ta)
ega bo‘lishdi	ega bo‘l (2ta)	ega bo‘lmoq	ega bo‘l (2ta)
Taqillatdi	Taqilla	taqillamoq	Taq
Undami	Un	u	U
Keldilar	Kel	kelmoq	Kel
Uyda	Uy	uy	Uy
har birimiz	har bir (2ta)	har bir	har bir (2ta)

In Uzbek language, the ordinary structure of word composition follows the pattern: "Root + Derivative + Lexical Form + Syntactic Form." The order in the placement of grammatical elements is determined by the significance and grammatical characteristics of the context. This is organized in the following sequence:

1. Creating new lexical meanings.

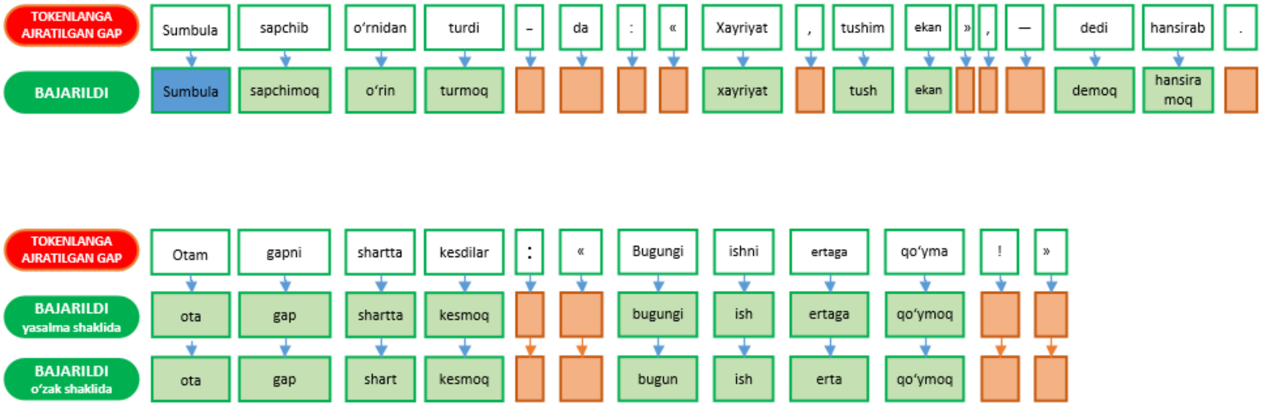
2. Influencing the lexical meaning.

3. Adding a component that does not affect the lexical meaning but is related to the word.

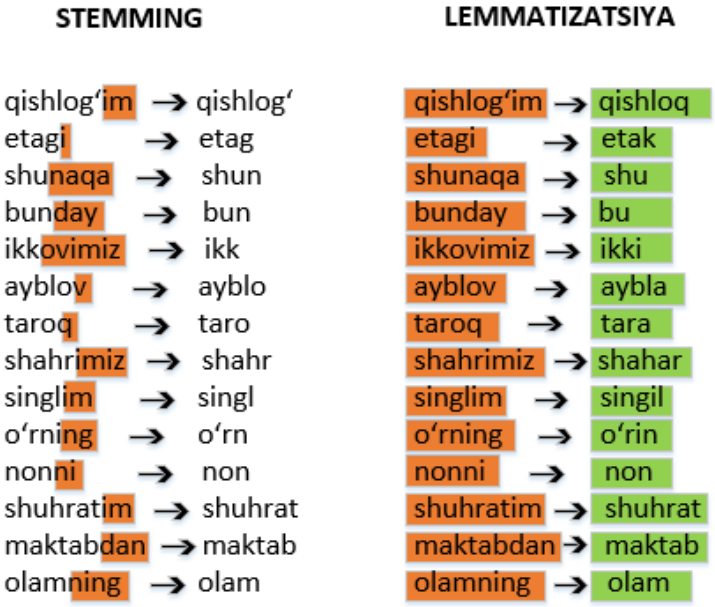
Occasionally, there are exceptional cases in the placement of additional components, such as in "opa-lar-im" or "opa-m-lar," and "ayt-di-ng-lar" or "ayt-di-lar-ing."

Suffixes in uzbek language are added after the root. Morphotactic rules in word formation analyze how morphemes and allomorphs are arranged in word composition. The lemmatization process is outlined as follows:

Copyright (c) 2023 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY).To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/



Picture 1. Examples of the lemmatization process

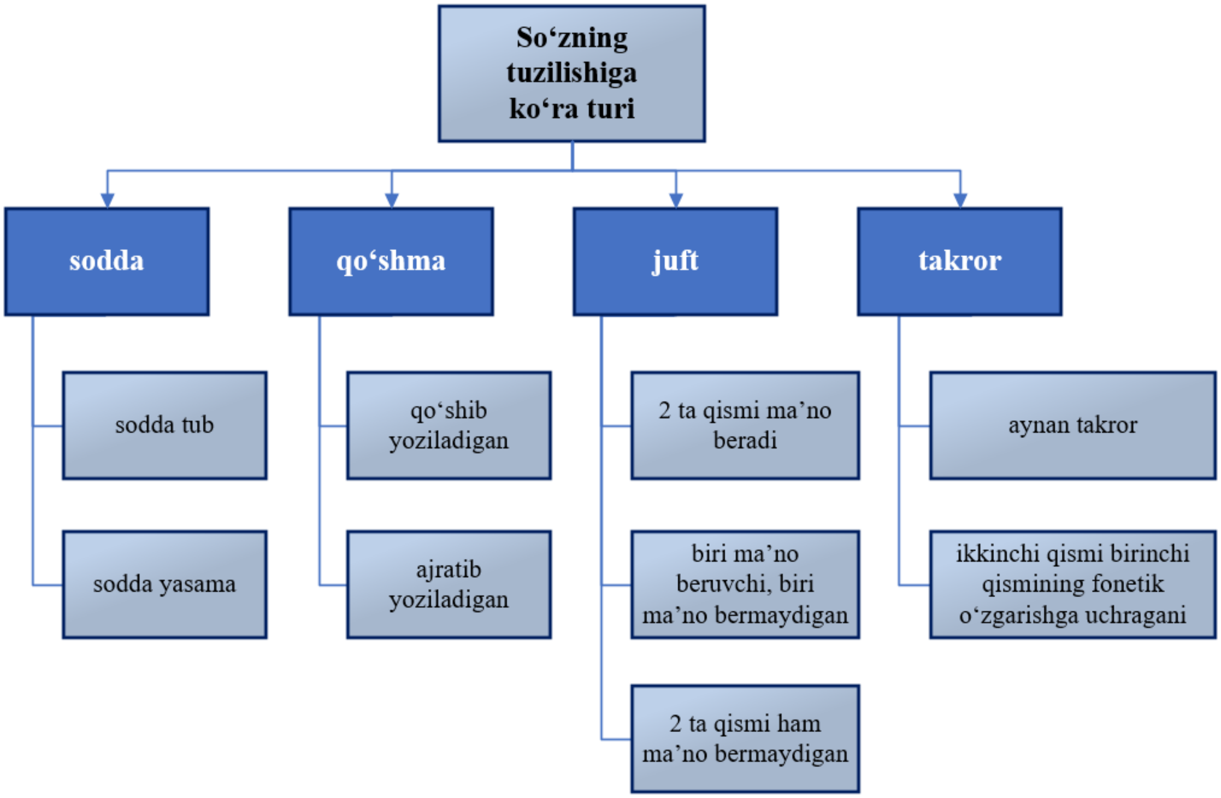


STEMMING		LEMMATIZATSIYA	
qishlog'im	→ qishlog'	qishlog'im	→ qishloq
etagi	→ etag	etagi	→ etak
shunaqa	→ shun	shunaqa	→ shu
bunday	→ bun	bunday	→ bu
ikkovimiz	→ ikk	ikkovimiz	→ ikki
ayblov	→ ayblo	ayblov	→ aybla
taroq	→ taro	taroq	→ tara
shahrimiz	→ shahr	shahrimiz	→ shahar
singlim	→ singl	singlim	→ singil
o'rning	→ o'rn	o'rning	→ o'rin
nonni	→ non	nonni	→ non
shuhratim	→ shuhrat	shuhratim	→ shuhrat
maktabdan	→ maktab	maktabdan	→ maktab
olamning	→ olam	olamning	→ olam

Picture 2. The difference between stemming and lemmatization

The topic discusses the necessity of attention to the aspects of stemming and lemmatization. Although in most cases, the results of stemming and lemmatization appear similar in the Uzbek language, they represent distinct processes: stemming involves cutting off additional affixes, while lemmatization identifies the base form in the dictionary. The text also provides examples illustrating the analysis of stemming and lemmatization processes in the structure of Uzbek words:

Copyright (c) 2023 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY).To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/


So‘zning tuzilishiga ko‘ra turi
sodda
qo‘shma
juft
takror
sodda tub
sodda yasama
qo‘shib yoziladigan
ajratib yoziladigan
2 ta qismi ma’no beradi
biri ma’no beruvchi, biri ma’no bermaydigan
2 ta qismi ham ma’no bermaydigan
aynan takror
ikkinchi qismi birinchi qismining fonetik o‘zgarishga uchragani

Picture 3. Classification of words in uzbek language according to their structure

Table 2. Example of simple and derived words

Simple words	Derived words
Lola	o‘qituvchi
maktab	ilmli
ona	vatandosh
matematika	tinchlik
ildiz	haydovchi

Simple words and derived words

Simple words and derived words are usually equivalent to a single lemma. To identify the lemma of these words, it is sufficient to eliminate inflectional suffixes.

Compound words

In determining the lemma of compound words, the further rules are followed:

- The lemma of compound words that are written as one word is determined by identifying suffixes which are used only for forming the shape of the words by cutting them off.
- In lemmatizing compound words which are written seperately, the dictionary is relied upon.

Copyright (c) 2023 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY).To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

- Proper nouns which consist of several words are written separately, they are NER: Katta Farg‘ona, Yangi O‘zbekiston, Birlash Millatlar Tashkiloti, Jahon Sog‘liqni Saqlash tashkiloti, Milliy Tiklanish partiyasi, Soliqni saqlash vazirligi.

Table 1. *Classification of compound words according to form*

Inseparable compounds	Separable compounds
Compound noun: *boltayutar, xontaxta, belkurak, xonqizi* **Compund adjective:** *kitobsevar, uchburchak, sho‘rtumshuq, balandparvoz, tezpishar, ertapishar, cho‘rtkesar*	**Compund verb:** *javob bermoq, ishlab chiqmoq, olib kelmoq, taklif etmoq* **Compound adverb:** *bir yo‘la, bir muncha, bir talay* **Compound pronoun:** *ana shu, mana bu, har bir*

There are several ideoms in uzbek language which are the same with compound words according to their structure. According to the composition of idioms, the following structures are present: :

- W1 + W2
- W1 + W2 + W3
- W1 + W2 + W3 + W4

Among these, ideoms with the composition W1 + W2 have a structural similarity with compound verbs. Verbs in the form of Ot+fe’l (*yaxshi ko‘rmoq, ko‘zda tutmoq, quloq solmoq, jon bermoq, ko‘zini uzmoq*) are equivalent to a single lemma but are not lemmas themselves; they are treated as phraseological unity.

Hyphenated words

In identifying the lemmas of hyphenated words, the further rules are followed:

- **Both parts have meaning:** achchiq-chuchuk. In this case, the stem is two, the lemma is one, and the dictionary is based on it.
- **One has meaning, the other does not**: kalta-kulta. In this case, the stem is only the meaningful part, but the lemma encompasses both parts. The dictionary is still based on it.
- **Neither part has meaning:** g‘idi-bidi, jiz-biz. If both parts of hyphenated words do not have meaning, the word itself becomes the stem, form, and lemma.

Reduplicative Words

In determining the lemmatization of repetitive words, the following rules are followed:

- **Exact repetition:** katta-katta, tez-tez, baland-baland.
- **The second part undergoes phonetic changes compared to the first part:** tuz-puz, non-pon.

Lemmatization of reduplicative words is also based on dictionary information. As seen above, it relies on dictionary information in lemmatization. Only in identifying the stem, shape-forming (syntactic and lexical) suffixes are omitted.

When one part of a reduplicative word comes separately, the POS tag is different from the original form; when the repetitive part returns, the POS tag belongs to another category:

Copyright (c) 2023 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY).To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

In using "dedi" alone, it is a verb, but when the reduplicative word "dedi-dedi" occurs, it belongs to the noun category.

However, this is not a typical case: tez ravish, tez-tez ravish; katta sifat, katta-katta sifat.

Conclusion

In information systems possessing a database of language corpora and texts, initial stages such as stemming or lemmatization are utilized to enhance the precision of information retrieval tasks. Both stemming and lemmatization processes play crucial roles in the information acquisition phase, where lemmatization aids search systems in identifying key words and reducing the size of index files. This article presents methods for lemmatizing various structured words in the Uzbek language, categorizing them into grammatical rules for simple, compound, derivative, hyphenated words and reduplicative words. Examples for each group were provided to illustrate the identification of word lemmas. The rules developed by the authors were applied to the Uzbek language morphological analyzer, resulting in a 97.8% accuracy rate.

References:

1. B. Elov, Sh. Hamroyeva, D. Elova. (2022). Methods for creating a morphological analyzer, Uzbekistan: language and culture (Practical philology), 2022, 5(1).
2. B. B. Elov, Sh. M. Hamroyeva, O. X. Abdullayeva, Z. Y. Xusainova, N. U. Xudayberganov. (2023). POS tagging and stemming in Uzbek, Turkic, and Uyghur languages, Uzbekistan: language and culture (computer linguistics), 2023, 1(6).
3. Ivan Boban, Alen Doko, Sven Gotovac. Sentence retrieval using Stemming and Lemmatization with Different Length of the Queries. Advances in Science, Technology, and Engineering Systems Journal Vol. 5, 2020, No. 3, 349-354.
4. Kharis, M., Laksono, K., Suhartono, Ridwan, A., Mintowati, & Yuniseffendri. (2022). Tokenization and Lemmatization on German Learning Textbook Level A1 of CEFR Standard. Journal of Higher Education Theory and Practice, 22(1). https://doi.org/10.33423/jhetp.v22i1.4971.
5. Uzbek language morphological analyzer - http://uznatcorpora.uz/.
6. B. Elov, Sh. Hamroyeva, X. Axmedova. (2022). Methods for creating a morphological analyzer, 14th International Conference on Intelligent Human-Computer Interaction, IHCI 2022, 19-23 October 2022, Tashkent.
7. Elov B., Alayev R., Xusainova Z. Implementation of stemming in Uzbek language: a hybrid statistical approach. 2023.

Copyright (c) 2023 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY).To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/