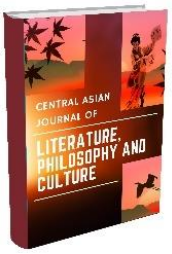

CENTRAL ASIAN JOURNAL OF LITERATURE, PHILOSOPHY AND CULTURE

eISSN: 2660-6828 | Volume: 04 Issue: 06 June 2023
https://cajlpc.centralasianstudies.org

The Problem of Words Undergoing Sound Changes in Uzbek Stemmers

Botir Elov
Doctor of philosophy of technical sciences (PhD), associate professor
Alisher Navo'i Tashkent State University of Uzbek Language and Literature

Zilola Xusainova
PhD student, Alisher Navo'i Tashkent State University of Uzbek Language and Literature

Habiba Berdiyeva
Master of degree, Alisher Navo'i Tashkent State University of Uzbek Language and Literature

Received 4th Apr 2023, Accepted 5th May 2023, Online 15th June 2023

ANNOTATION

Stemming is one of the most common initial data processing steps that can be performed on almost all Natural Language Processing (NLP) projects. In the process of Stemming, it is carried out to remove some part of the word or shorten the word to its root. Several stemming algorithms are used to decide how to cut a word. In determining the stem of Uzbek words, problems such as homonymy of root and suffix with one root, sound changes when the suffix is added to the words, stemming of neologisms and NERs can occur. This article presents models for solving the problem of the occurrence of sound changes in words in the process of performing stemming in the texts of the Uzbek language Corpus.

KEYWORDS: Natural Language Processing, NLP, root, stem, sound change, POS tagging, morphological analyzer.

INTRODUCTION. In Uzbek language, a word is formed by combining the vowels and suffixes (affixes). Analysis of both phonetic and morphological changes is an important task as phonetic harmony and disharmony occur when Affixes are added to the root. When solving tasks of NLP, word forms have to be reduced to the root (stemming). Removing all flective affixes from a word and lemmatizing the rest of the word is considered one of the important tasks of Natural Language Processing (NLP), and this process is referred to as stemming. Stemming process is important in information search (IR, Information Retrieval) Systems [1].

In Uzbek language, where the smallest meaningful part of a word is defined as a root, stem is the largest part that gives meaning to a word. Therefore, we can say that the word consists of two parts: stem, which explicates meaning, and suffixes.

Stem is the part that is formed from the excision of the suffixes of the word form, and may not mean in some cases. Also, the stem doesn't exactly match or match the morphological root of the word.

Stemming is the task of shortening to its root by removing derivational and flective suffixes added to the word. The stemming process can be seen as a "rough" heuristic process that simply cuts off the suffixes of words. According to the authors, unlike lemmatization, the stemming process does not use vocabulary or morphological analysis [2]. The root formed by stemming does not necessarily resemble the actual word or its morphological root. The aim of the stemming process is to shorten similar words to the same root.

Words in Uzbek language are formed by adding some suffixes (affixes) to the root. In some cases, phonetic changes can occur in the word, and this is reflected directly in the text. A root itself may also be a word that expresses a specific meaning of a word. While affixes play an important role in a sentence, they do not have an independent meaning. Affixes are classified into derivational suffixes and inflectional suffixes [3]. Inflectional suffixes change only the grammatical function of the word. By adding derivational suffixes to the root, a semantic change in the word can occur. Inflectional suffixes produce syntactic changes in the word. A derivational suffix is attached to the root first, followed by an inflectional suffix [4].

The number of suffixes that can be added to a word and their multiple compounds make the process of specifying a root a complex problem in agglutinative languages. Because in most agglutinative languages, the combination of suffixes produces complex word forms. Indicators for the number of new derivational and inflectional suffixes are also calculated differently.

To determine the stem of words in Uzbek, the root and all kinds of suffixes that attach to it are determined. Traditional stemming algorithms are based on suffixes and some morphological rules, and uncertainty in stem may result from the stemming process. In the process of stemming, all types of suffixes in the word are usually removed. But when stemming is performed in this way, a wrong result can be obtained in some cases [5].

The following problems may arise when determining the stem of words in Uzbek language:

- *homonymy of root and suffix with one stem,*
- *sound changes in the word;*
- *stemming of neologisms and NERs.*

In this research work, problems with sound change of a word and methods for solving them are presented when determining the stem of words in Uzbek language.

Sound changes in the word

As a phoneme is realized in speech, its variant manifests differently due to the colloquial conditions (exposure to adjacent sounds or suffixes). As a result, the sound acquires properties that are not in the essence of the word. Some sound adapts to the adjacent sound, some of which intensify and alternate to another sound. The change that occurs in speech is called a combinatorial-positional change in sound[6]. We come across such changes a lot in the process of speech activity.

This situation occurs in the way that a voicing consonant changes to a devoicing one, a plosive consonant changes to a nasal consonant, or one wide vowel changes to another wide vowel or some sounds are dropped or increased as a result of the addition of auxiliary morphemes to the word composition. Significantly, some

of the changes noted above are reflected in both pronunciation and writing, while some occur only in the oral speech process and are not reflected in writing.

Speech sounds that are part of a syllable, word, phrase affect each other, resulting in sound changes. The sound change that occurs in the speech stream is called phonetic process[7]. Therefore, any sound change occurs as a result of the influence of speech sounds on each other. Such a phonetic process and a detailed study of phonetic phenomena open a wide path to our deeper study of the content and essence of Phonetics and Phonology. The addition of inflectional suffixes to the end of the root may result in phonetic changes in the word (insertion, deletion, phonetic harmony, and assimilation) in some cases [4].

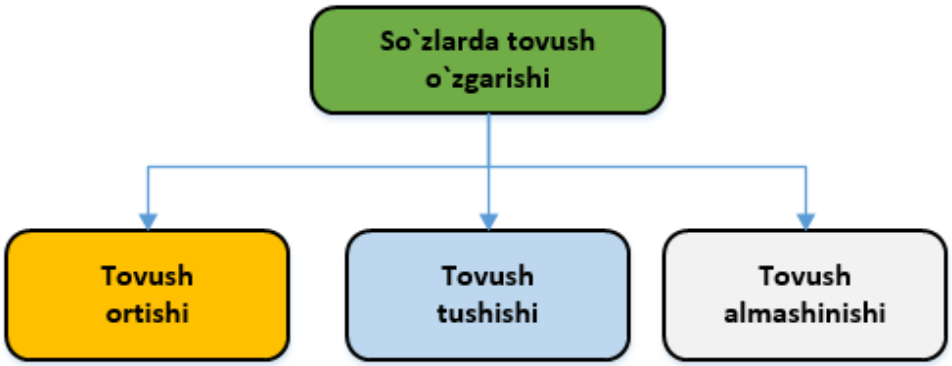

Figure 1. Sound changes in the word

In Uzbek languae, three different phonetic changes can be made in a word, such as assimilation, dissimilation and metathesis (Table 1).

Table 1. Assimilation, dissimilation and metathesis in Uzbek language

assimilation	dissimilation	metathesis
obro‘y+imiz achch+iq	me+ning singl+isi	boshlig‘+i tarog‘+ini
shun+day	ko‘ngl+im	san+aydi
parvoy+im un+ga	bo‘yn+i olt+ovlon	yurag+im qulog‘+ing

To solve the problem of sound change in stem detection, the boundaries of the root and suffixes are determined in the first step, while lemmatization is carried out in the second step. As a result of lemmatization, the error generated stems are changed to root, which is available in the dictionary.

Orthographic rules

When certain suffixes are added to some roots, in the root or suffixes, phenomena associated with the alternation of one sound to another are also observed. These changes are reflected not only in pronunciation, but also in writing.

1. Metathesis in words

Metathesis is considered to apply equally to both vowels and consonant sounds.

1.1.when the suffixes**-v, -q, -qi** are added to verbs ending in vowel **a**, vowel **a** is pronounced **o** and is written as:

root	suffix	word form	stem
sayl**a**	-v	sayl**o**v	sayl**o**
tar**a**	-q	tar**o**q	tar**o**
sayr**a**	-qi	sayr**o**qi	sayr**o**

1.2. the suffixes**-v, -q** are added to most verbs ending in vowel **i**. this vowel is pronounced **u** and is written as

root	suffix	word form	stem
qaz**i**	-v	qaz**u**v	qaz**u**
sov**i**	-q	sov**u**q	sov**u**
o‘q**i**	-v	o‘q**u**v	o‘q**u**

Note. But when the suffix **-q** is added to certain verbs that end in vowel **i**, this vowel **i** is pronounced and written like this[1]:

root,stem	suffix	word form
og‘r**i**	-q	og‘ri**q**
qav**i**	-q	qavi**q**

1.3. When the possessive suffix is added to the multi-syllable and mono-syllable words ending with consonants **k, q**, the consonant **k** becomes the consonant **g**, the consonant **q** becomes the consonant **g**, and is written as.

Root	Suffixes	Word form	Stem
yura**k**	{-im,-ing,-si,-i}	yura**g**im	yura**g**
qulo**q**	{-im,-ing,-si,-i}	qulo**g‘**im	qulo**g‘**
yo‘**q**	{-im,-ing,-si,-i}	yo‘**g‘**im	yo‘**g‘**

Note. But when the possessive suffix is added to the multi-syllable and mono-syllable words, the sound **k, q** is originally pronounced and written as

root, stem	suffix	Word form
zav**q**	{-im,-ing,-si,-i}	zav**q**i
par**k**	{-im,-ing,-si,-i}	par**k**i
no**k**	{-im,-ing,-si,-i}	no**k**i

1.4. Sound change is observed by the addition of derivational suffixes such as **-a, -ay, -la** to some words[2]:

root	suffix	Word form	stem
ot	-a	**a**ta	**a**t
ong	-la	**a**ngla	**a**ng
sariq	-ay	sar**g‘**ay	sar**g‘**

[1] https://github.com/KhZilola/Python-Codes/blob/main/Tovush%20almashishi.docx
[2] https://github.com/KhZilola/Python-Codes/blob/main/Istisnolar.docx

2. Assimilation in words

There are two different manifestations of assimilation that are reflected only in pronunciation and expressed both in pronunciation and in writing. The phenomenon that manifests only in the process of oral speech occurs mainly within the framework of loanwords: *rus→o'ris, shkaf→ishkof, stakan→istakan, traktor→tiraktor*.

The current Uzbek literary language textbook states that the changes observed in the above examples are examples of the phenomenon of prosthesis[9]. This phenomenon is explained by the fact that in Turkic, including Uzbek, the juxtaposition of consonant sounds at the beginning of a word is very rare, therefore, in order to achieve ease of pronunciation, one vowel is first increased and pronounced before them. There are several manifestations of assimilation in Uzbek linguistics that have been thoroughly researched by linguistic scholars.

Table 2. Rules for the assimilation in the word in Uzbek language

Type of assimilation	How it happens?	Examples
Prosthesis	It is a phenomenon associated with the addition and pronunciation of a single vowel at the beginning of a word, before the sound of sonor [r], the vowel [o'] is increased	*ro'mol→[o'ramol], ro'za→[o'raza], rais→[o'rais], rang→[o'rang]*
Epenthesis	When it comes to two consonant rows at the beginning, middle and end of the word,among them, the vowel [i], sometimes [u] and [a] are increased and pronounced	*fikr→[fikir], hukm→[hukum], doklad→[dakalad], klass→[kilass]*

2.1.When the suffixes **-da, -dan, -day, -dagi, -ga, -gacha, -cha are added to pronouns including u, bu, shu, o'sha, the sound n is added and written as**

root	suffix	Word form	stem
u	-ga	u**n**ga	un
bu	-ga	bu**n**ga	bun
shu	-dagi	shu**n**dagi	shun
shu	-ga	shu**n**ga	shun

2.2.When first-person, second-person possessive suffixes are added to words **parvo, obro', mavqe, mavzu, avzo, sound y is added and is written as**

root	suffix	Word form	stem
parvo	-im	parvo**y**im	parvoy
obro'	-im	obro'**y**im	obro'y
mavqe	-im	mavqe**y**im	mavqey
mavzu	-im	mavzu**y**im	mavzuy
avzo	-im	avzo**y**im	avzoy

Note: Third-person possessive suffix are added to wods **xudo, mavzu** as **-si** and no assimilation is observed.

We can find some information about prosthesis in other literature. Adding sounds to the initial part of a word is called a prosthesis. In Turkic languages, including Uzbek, words entered from Russian are pronounced with the addition of vowels *[i], [u],* before the consonant compounds that come at the beginning. *stansiya→[istansiya], stol→[ustol].* But this phenomenon is decreasing[7].

In some cases, [I] vowels can also be assimilated when two consonant rows come in at the beginning of a word, which are plosive and nasal. *shkaf→[ishkaf], spravka→[ispravka], stol→[istol], stul→[istul], shtraf→[ishtraf]*

3. Dissimilation in words

Speech sounds are the material that makes up speech. Speech saves this material in its construction. This saving of speech sounds in the speech is seen as its fall, that is, the phenomenon of dieresis [10].

3.1.When the possessive suffix is added to some words such as **o‘rin, qorin, burun, o‘g‘il, bo‘yin, ko‘ngil**, the vowel in the second syllable is not pronounced and not written[3]

root	suffix	Word form	stem
o‘rin	-i	o‘rn**i** (o‘r~~i~~ni)	o‘rn
qorin	-i	qorn**i** (qor~~i~~ni)	qorn
burun	-i	burn**i** (bur~~u~~ni)	burn
o‘g‘il	-i	o‘g‘l**i** (o‘g‘~~i~~li)	o‘g‘l
bo‘yin	-i	bo‘yn**i** (bo‘y~~i~~ni)	bo‘yn
ko‘ngil	-i	ko‘ngl**i** (ko‘ng~~i~~li)	ko‘ngl

3.2.When the suffix**-il** is added to verbs such as **qayir, ayir**, the vowel in the second syllable is not pronounced and not written

root	suffix	Word form	stem
qayir	-il	qayr**il** (qay~~i~~ril)	qayr
ayir	-il	ayr**il** (ay~~i~~ril)	ayr

3.3.The vowel in the second syllable is not pronounced and not written when adding the suffixes **-ov,- ala, -ovlon** to the words **ikki, olti, yeti**

root	suffix	Word form	stem
ikki	-ov	ikk**ov**(ikk~~i~~ov)	ikk
olti	-ala	olt**ala** (olt~~i~~ala)	olt
yetti	-ovlon	yett**ovlon**(yett~~i~~ovlon)	yett

3.4.When the suffix –a, –ay, is added to some words, the vowel in the second syllable is not pronounced and not written

root	suffix	Word form	stem
o‘yin	-a	o‘yn**a** (o‘y~~i~~na)	o‘yn
ulug‘	-ay	ulg‘**ay** (ul~~u~~g‘ay)	ulg‘
sariq	-ay	sarg‘**ay** (sar~~iq~~ay)	sarg‘

Note. The phenomenon of metathesis is observed in the word **sarg’ay** in the form **sariq+ay=sarg’ay**

[3] https://github.com/KhZilola/Python-Codes/blob/main/Tovush%20tushishi.docx

3.5.When the suffixes**-ni, -ning, -niki** are added to the pronouns of **men, sen,** the consonant **n** in the suffix is not pronounced and not written

root	suffix	Word form	stem
men	-ni	men**i** (men~~n~~i)	men
sen	-ning	sen**ing** (sen~~n~~ing)	sen
men	-niki	men**iki** (men~~n~~iki)	men

In different cases, different changes occur in the pronunciation of sounds. N. S. Trubetskoy shows that "... practically it is impossible to pronounce exactly one sound in one position even several times clearly and in one type " [11]. That is, the speaker pronounces his words in different tones, both in different situations and in the same situations. Therefore, each time the sounds of speech are pronounced in a different tone, another- with different pitch and quality indicators. A single sound will never repeat exactly as it is. The deeper the positional variations of sounds are studied in speech, the wider the types of sounds can be identified and a broader picture can be drawn in this regard. In the "Stemming" part of the software of the morphological analyzer of the Uzbek language[4], word forms are analyzed and some examples related to the phenomenon of sound change are cited below.

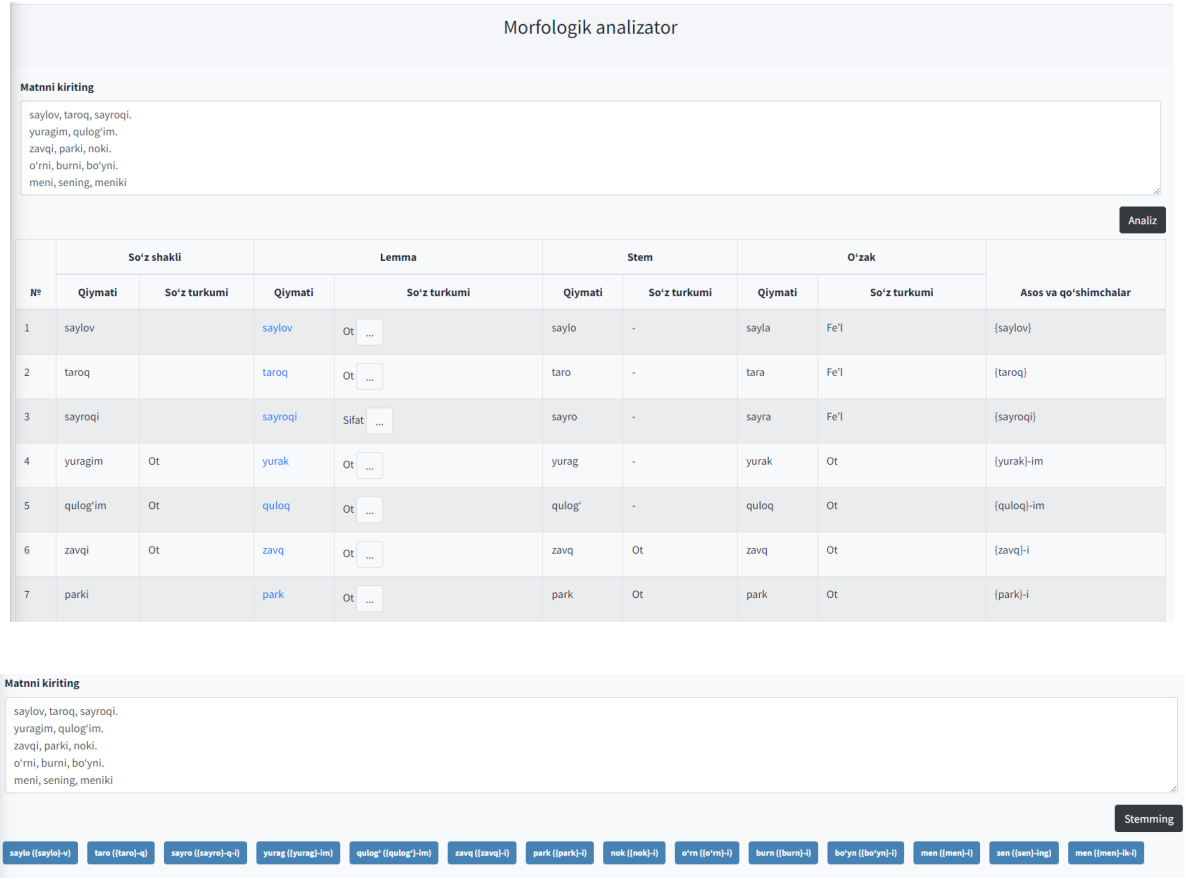

Figure 2. Morphological analyzer of Uzbek language (stemming)

[4] http://uznatcorpara.uz/uz/Stemmer

Conclusion

In this article, the issue of sound change in a word in the process of performing stemming in the texts of the Uzbek language corpus was subjected to analysis. In the article, three different types of phonetic changes in certain words of Uzbek, such as assimilation, dissimilation and metathesis, were analyzed on the basis of grammatical rules and covered by the means of examples. To solve the problem of sound change in stem detection, the boundaries of the root and suffixes are determined in the first step, and lemmatization is carried out in the second step. As a result of lemmatization, the erroneously generated stems are changed to an existing root in the dictionary. The methods presented in the article were applied to the morphological analyzer of the Uzbek language developed by B.Elov, R. Alaev, Sh. Khamroyeva and Z. Khusainova.

References

1. Elov B.B., Khamroyeva Sh.M., Abdullayeva O.X., Khusainova Z.Y., Xudayberganov N.U. POS tagging and stemming in Uzbek, Turkish and Uyghur languages. Uzbekistan: language and culture (computer linguistics), 2023/1(6), 41-60 pp.
2. B.B.Elov, Sh.M.Khamroeva, Z.Y.Khusainova. Pipeline conveyer of NLP (natural language processing). Descendants of Muhammad al-Khwarazmi. Scientific-practical and information – analytical Journal, 1 (23) / 2023, 181-192 pp.
3. Sharma, A., Kumar, R., & Mansotra, V. (2016). Proposed Stemming Algorithm for Hindi Information Retrieval. International Journal of Innovative Research in Computer and Communication Engineering (An ISO Certified Organization), 3297(6). https://doi.org/10.15680/IJIRCCE.2016
4. Hajiev A. Word making in Uzbek language. - Tashkent, 2005
5. Paice, C. D. (1990). Another Stemmer. *ACM SIGIR Forum*, *24*(3). https://doi.org/10.1145/101306.101310
6. Sayfullaeva R., Mengliyev B. et al. Current Uzbek literary language. - Tashkent: 2009, 65 p
7. Sadigov A. et al., An introduction to linguistics. – Tashkent. O'qituvchi, 1981. 47-51 pp.
8. Practicum of Uzbek language. Part 1. - Tashkent: 2005, 12 p
9. Sayfullayeva R., Mengliyev B. et al. Modern Uzbek literary language. -Tashkent: 2009, p 671
10. Mirtojiyev M. Phonetics of the Uzbek language. -Tashkent: "Fan", 2013. 307 p
11. N.S. Trubetskoy. Fundamentals of Phonology, M., 1960. 48-49 pp