

Computational Model of Morphology and Stemming of Uzbek Words on Complete Set of Inflectional Endings

Nargiza Gabdullina, Nazerke Karipbayeva, Nilufar
Abdurakhmonova, Ualsher Tukeyev

Plan:

1. Introduction

2. Computational Model of the Uzbek Language
Morphology on the Complete Set of Inflectional
Endings

 2.1. Inferring of Inflectional Endings for
Nominal Base Words

 2.2. Inferring of Inflectional Endings for Verb
Base Words

3. Experiments and Results

4. Conclusions and Future Works

Introduction

The Uzbek language belongs to the group of Turkic-speaking languages and is one of the low-resource languages. In this regard, it is currently important to increase and expand the language and electronic resources in the Uzbek language. However, since the Uzbek language belongs to the group of agglutinative languages, in this language each grammatical meaning is expressed by separate affixes. Therefore, when constructing natural language processing tasks, such as stemming, segmentation and morphological analysis, a complete set of endings is required along with the stem and stop words of the Uzbek language. The article contains a full set of Uzbek endings, a dictionary of stem and stop words. The collection of endings was carried out for two main parts of speech, that is, for the noun and the verb. The dictionary of verb endings includes all possible combinations of tenses, voices, moods, and participles.

Inflectional Endings for Nominal Base Words

The set of endings to the nominal bases of words of the Uzbek language has four types:

- plural suffixes (denoted by K),
- possessive suffixes (denoted by T),
- case suffixes (denoted by C),
- personal suffixes (denoted by J).

Inflectional Endings for Nominal Base Words

Ending type	Endings	Number of endings
K	-lar	1
T	-im, m,-ing,-ng, -i, -si, -imiz, -miz, ingiz, -ngiz, -niki	11
C	-ning, -ga, -ka, -qa, - ni, -dan, -da	7
J	-man, -san, -miz, -siz, -dir, -dirlar	6

Inflectional Endings for Nominal Base Words

Placements of two types of endings can be as follows:

KT, TC, CJ, JK

KC, TJ, CT, JT

KJ, TK, CK, JC.

Example	ending type K	ending type C	Number of endings
kitob-	-lar	-ning, -ga, -ni, -da, -dan	5

Number of endings of the KC (Plural-Case) endings placements.

Inflectional Endings for Nominal Base Words

The endings of the three types will be placed as follows:

KTC, KTJ, TCJ, TCK, CJK, CJT, JKT, JKC

KCJ, KCT, TJK, TJC, CTK, CTJ, JTK, JTC

KJT, KJC, TKC, TKJ, CKT, CKJ, JCK, JCT.

example	ending type K	ending type C	ending type J		Number of endings
			singular	plural	
ona-	-lar	-ga -da -dan	-man -san -dir	-miz -siz	C3) $1*5=5$ C5) $1*5=5$ C6) $1*5=5$ $3*5=15$

Number of endings of the KCJ (Plural-Case-Personal) endings placements.

The number of endings for nominal base words – 360.

Inflectional Endings for Verb Base Words

12 tenses in the Uzbek language, and all possible forms of the question and the negative were considered.

examples		affixes	1 person	2 person	2 person (respect)	3 person	Number of endings
After consonant	kel-	-ayotir	-man	-san	-siz	-	4
			-miz	-siz	-sizlar	-lar	3
After vowel	o'qi	-(y)yotir					4 3 7*2=14

The third method of forming a present continuous tense verb.

Inflectional Endings for Verb Base Words

12 tenses in the Uzbek language, and all possible forms of the question and the negative were considered.

examples		affixes	1 person	2 person	2 person (respect)	3 person	Number of endings
After consonant	kel-	-ayotir	-man	-san	-siz	-	4
			-miz	-siz	-sizlar	-lar	3
After vowel	o'qi	-(y)yotir					4 3 7*2=14

The third method of forming a present continuous tense verb.

The number of endings for verb base words – 1868.

The total number of endings for Uzbek – 2228.

Stopwords and stems

CLTICS - uzbek language — Блокнот

Файл Правка Формат Вид Справка

men
sen
biz
siz
u
ular
mening
sening
seniki
meniki
kim
nima
hech kim
hech nima
qanday
qanaqa
qaysi
qancha
necha
nechta
nechanchi
nega

stems — Блокнот

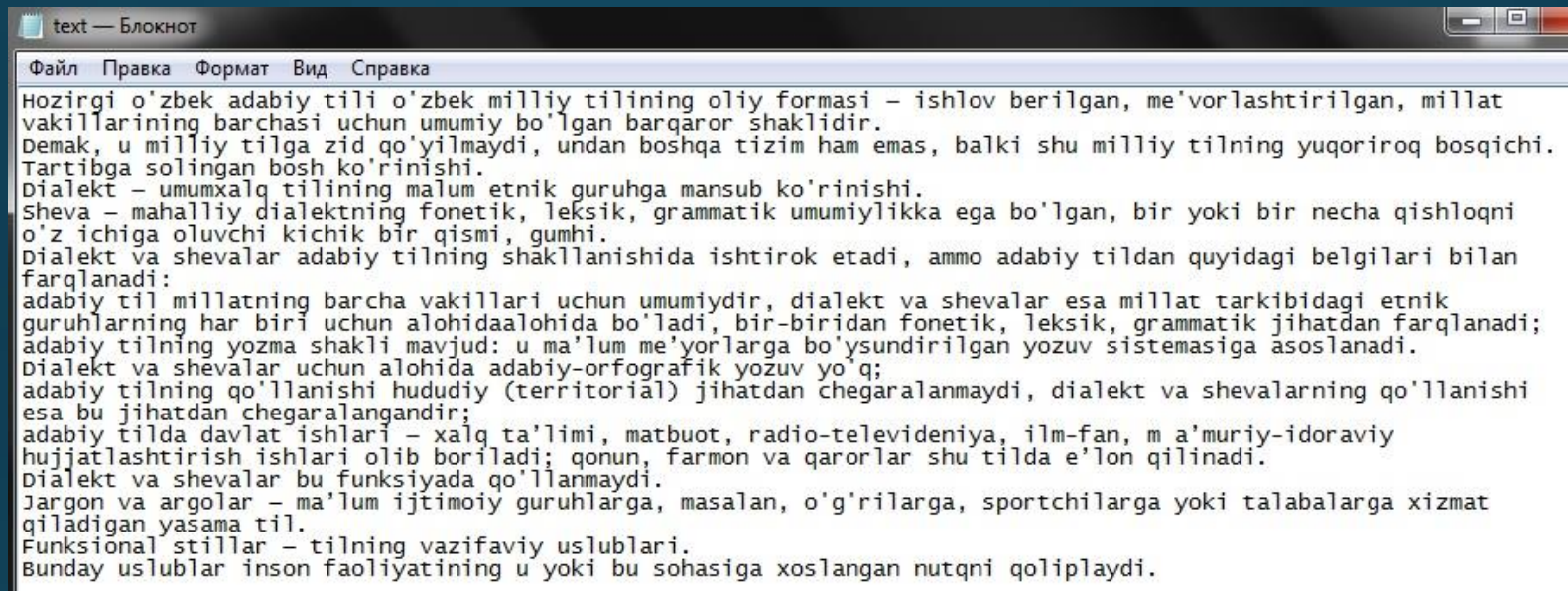
Файл Правка Формат Вид Справка

abad
abadiy
abadiya
abadiylash
abadiylik
abajur
abas
abbat
abbreviatura
abdol
aberratsion
aberratsiya
abgor
abgorlik
abira
abiturient
abjad
abjaqlamoq
abjirlik
ablah
ablahlik
ablahona
ablaq
abonement
abonent

	A	B
1	аффиксы	
2	lar	
3	im	
4	m	
5	ing	
6	ng	
7	i	
8	si	
9	imiz	
10	miz	
11	ingiz	
12	ngiz	
13	ning	
14	ga	
15	ka	

Experiments and Results

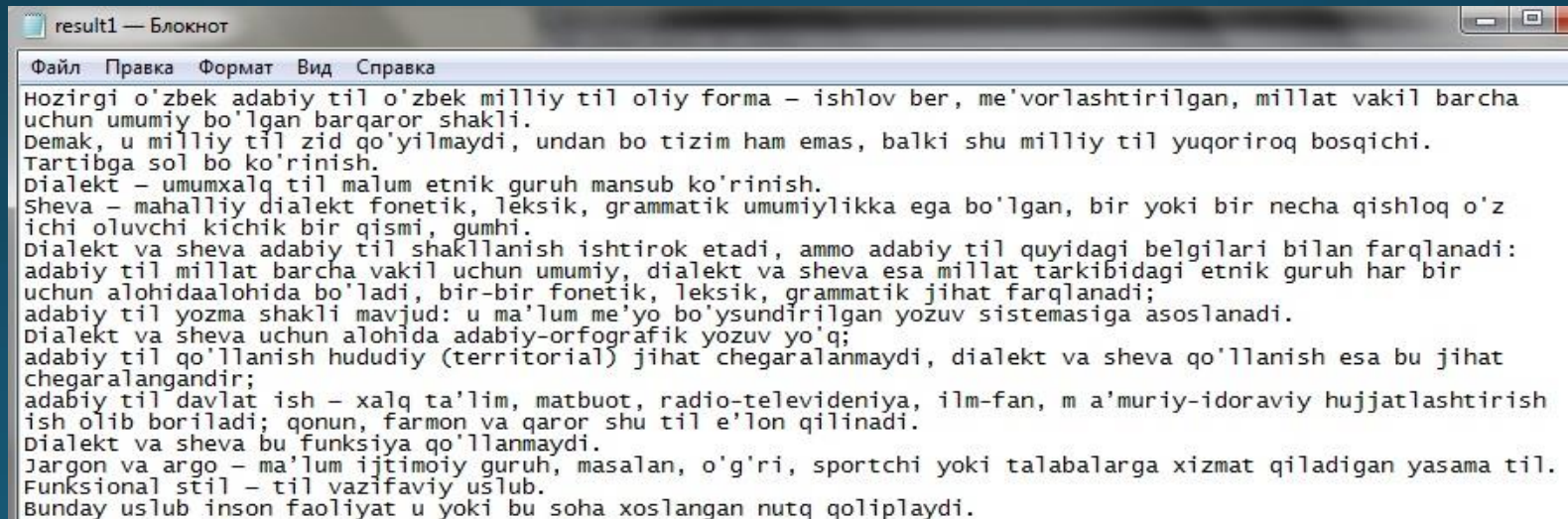
Text:



Файл Правка Формат Вид Справка

Hozirgi o'zbek adabiy tili o'zbek milliy tilining oliy formasi – ishlov berilgan, me'vorlashtirilgan, millat vakillarining barchasi uchun umumiy bo'lgan barqaror shaklidir.
Demak, u milliy tilga zid qo'yilmaydi, undan boshqa tizim ham emas, balki shu milliy tilning yuqoriroq bosqichi.
Tartibga solingan bosh ko'rinishi.
Dialekt – umumxalq tilining malum etnik guruhga mansub ko'rinishi.
Sheva – mahalliy dialektning fonetik, leksik, grammatik umumiylikka ega bo'lgan, bir yoki bir necha qishloqni o'z ichiga oluvchi kichik bir qismi, gumhi.
Dialekt va shevalar adabiy tilning shakllanishida ishtirok etadi, ammo adabiy tildan quyidagi belgilari bilan farqlanadi:
adabiy til millatning barcha vakillari uchun umumiydir, dialekt va shevalar esa millat tarkibidagi etnik guruhlarining har biri uchun alohidaalohida bo'ladi, bir-biridan fonetik, leksik, grammatik jihatdan farqlanadi;
adabiy tilning yozma shakli mavjud: u ma'lum me'yorga bo'ysundirilgan yozuv sistemasiga asoslanadi.
Dialekt va shevalar uchun alohida adabiy-orfografik yozuv yo'q;
adabiy tilning qo'llanishi hududiy (territorial) jihatdan chegaralanmaydi, dialekt va shevalarning qo'llanishi esa bu jihatdan chegaralangandir;
adabiy tilda davlat ishlari – xalq ta'limi, matbuot, radio-televideniya, ilm-fan, m a'muriy-idoraviy hujjatlashtirish ishlari olib boriladi; qonun, farmon va qarorlar shu tilda e'lon qilinadi.
Dialekt va shevalar bu funksiyada qo'llanmaydi.
Jargon va argolar – ma'lum ijtimoiy guruhlarga, masalan, o'g'rilarga, sportchilarga yoki talabalarga xizmat qiladigan yasama til.
Funksional stillar – tilning vazifaviy uslublari.
Bunday uslublilar inson faoliyatining u yoki bu sohasiga xoslangan nutqni qoliplaydi.

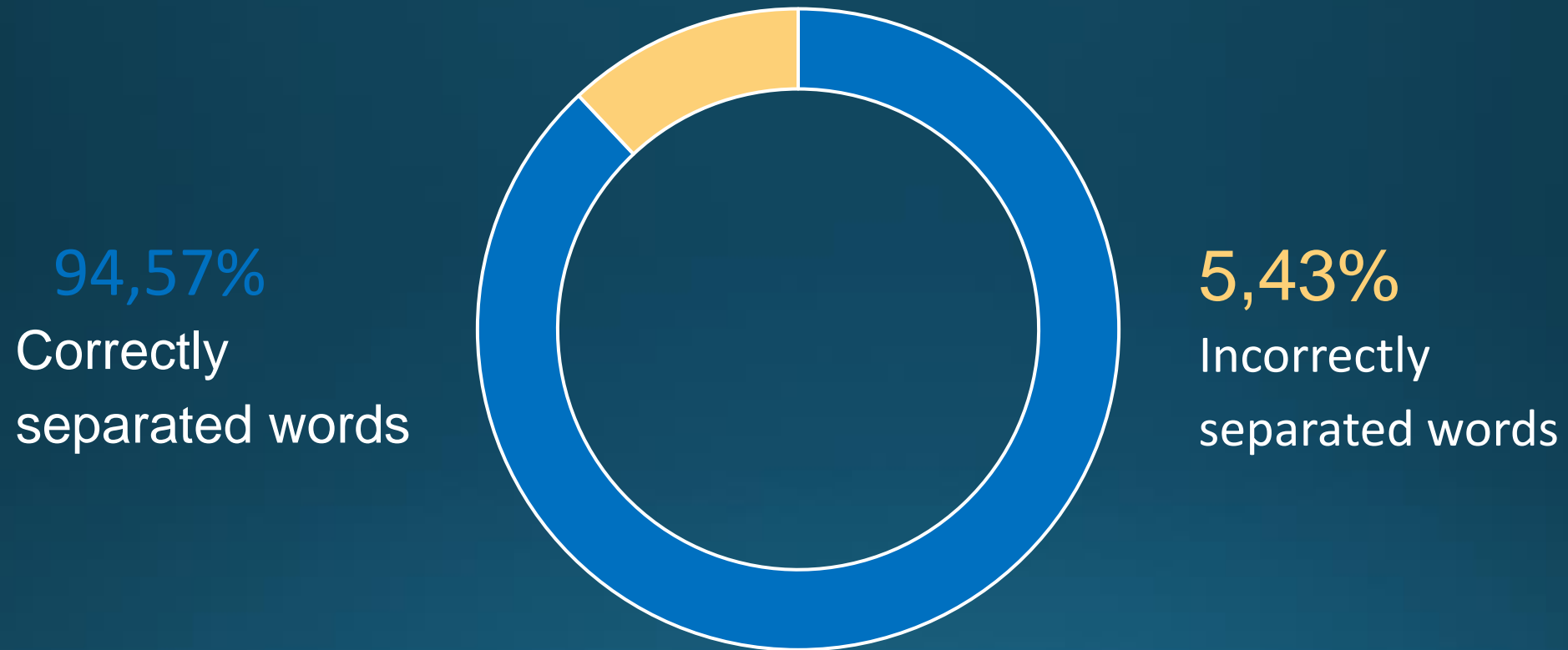
Stemming result:



Файл Правка Формат Вид Справка

Hozirgi o'zbek adabiy til o'zbek milliy til oliy forma – ishlov ber, me'vorlashtirilgan, millat vakil barcha uchun umumiy bo'lgan barqaror shakli.
Demak, u milliy til zid qo'yilmaydi, undan bo tizim ham emas, balki shu milliy til yuqoriroq bosqichi.
Tartibga sol bo ko'rinish.
Dialekt – umumxalq til malum etnik guruh mansub ko'rinish.
Sheva – mahalliy dialekt fonetik, leksik, grammatik umumiylikka ega bo'lgan, bir yoki bir necha qishloq o'z ichi oluvchi kichik bir qismi, gumhi.
Dialekt va sheva adabiy til shakllanish ishtirok etadi, ammo adabiy til quyidagi belgilari bilan farqlanadi:
adabiy til millat barcha vakil uchun umumiy, dialekt va sheva esa millat tarkibidagi etnik guruh har bir uchun alohidaalohida bo'ladi, bir-bir fonetik, leksik, grammatik jihat farqlanadi;
adabiy til yozma shakli mavjud: u ma'lum me'yo bo'ysundirilgan yozuv sistemasiga asoslanadi.
Dialekt va sheva uchun alohida adabiy-orfografik yozuv yo'q;
adabiy til qo'llanish hududiy (territorial) jihat chegaralanmaydi, dialekt va sheva qo'llanish esa bu jihat chegaralangandir;
adabiy til davlat ish – xalq ta'lim, matbuot, radio-televideniya, ilm-fan, m a'muriy-idoraviy hujjatlashtirish ish olib boriladi; qonun, farmon va qaror shu til e'lon qilinadi.
Dialekt va sheva bu funksiya qo'llanmaydi.
Jargon va argo – ma'lum ijtimoiy guruh, masalan, o'g'ri, sportchi yoki talabalarga xizmat qiladigan yasama til.
Funksional stil – til vazifaviy uslub.
Bunday uslub inson faoliyat u yoki bu soha xoslangan nutq qoliplaydi.

Experiments and Results



The experiment included 55 sentences of 626 words. The stemming algorithm correctly separated 592 words.

Experiments and Results

Inflectional word	Output	Expected output
kompyuterlarni	kompyuter	kompyuter
bloki	blok	blok
qanday	qand	qanday
disklardagi	disklardagi	disk
qiladi	qil	qil
texnikadan	texnika	texnika
yarat	yara	yarat
saqlanib	saqlan	saqlan

Possible results of applying the stemming algorithm based on the CSE morphological model to the text in the Uzbek language

Conclusion

- The article takes several new resources of the Uzbek language. This is: a complete set of endings of the Uzbek language, a dictionary of stem and stop words. And the result of the experiment using the accumulated linguistic resources showed an accuracy of 94.5%.
- In the future, considering the word-forming suffixes of nouns in the Uzbek language and their combinations with case, plural, possessive affixes, it is possible to increase the percentage of accuracy from this result for the better. And in the future, the dictionary of endings and stems of the Uzbek language will be used in the morphological analysis of Uzbek texts and text segmentation in neural machine translation.

Thank you for your attention!