Morphology Model and Stemming of Turkish Words on Complete Set of Inflectional Endings

Aitan Qamet, Ualsher Tukeyev

Al-Farabi Kazakh National University, Almaty, Kazakhstan

Faculty of Information Technology

qametaitan@gmail.com, ualsher.tukeyev@gmail.com,

Turklang 2021

The purpose of the work

Research and development of Morphology Model and Stemming of Turkish Words on Complete Set of Inflectional Endings

Tasks of the work

01

Description of the CSE (Complete Set of Endings) morphology model combinatorial approach

02

Review of methods and technologies for stemming of

Turkish Words

03

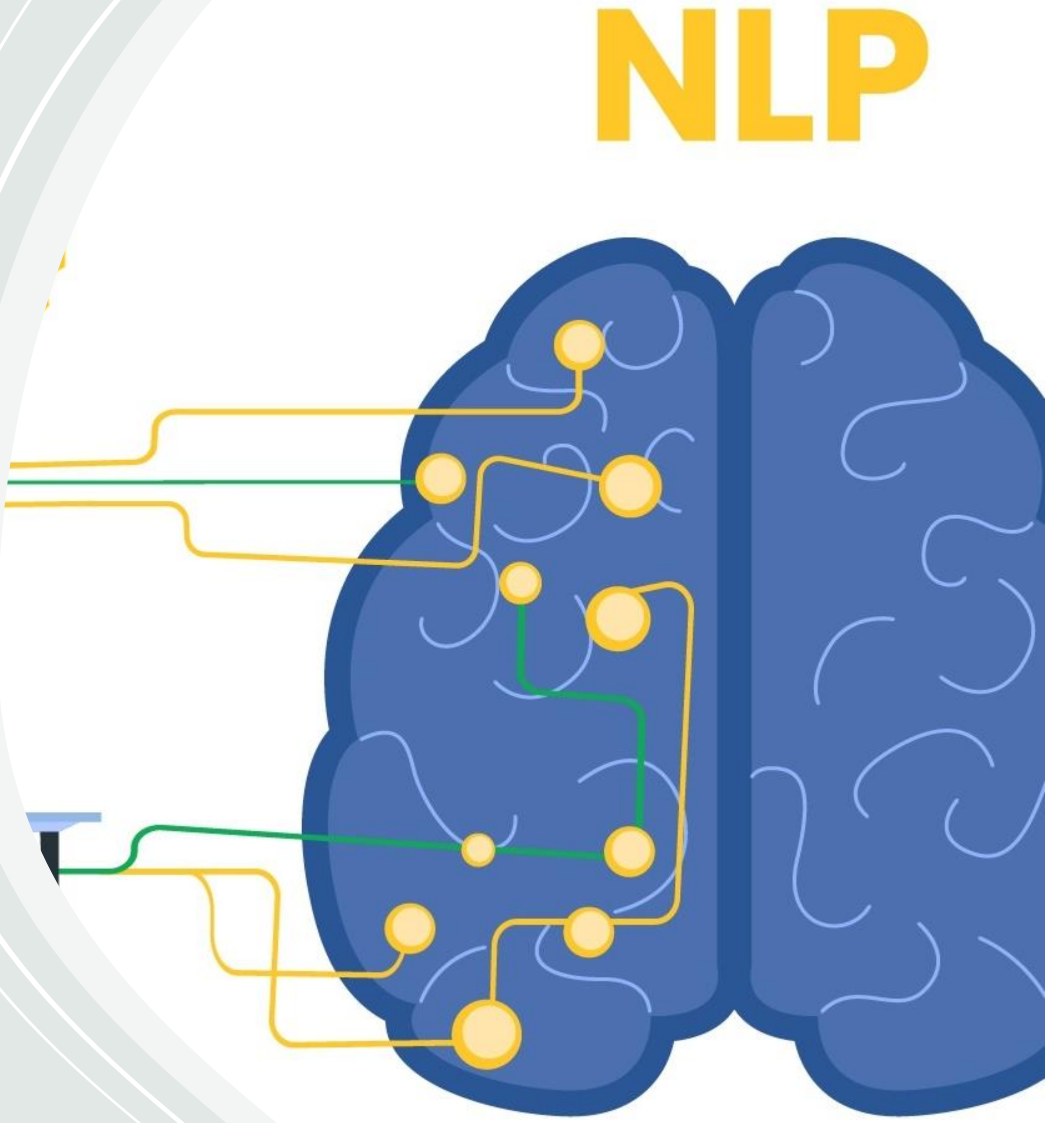
Design and development of a stemming of

Turkish words

Introduction

We can talk about the problem of the development of the NLP direction for the Turkic languages in general, the creation of convenient models and technologies for the development of NLP for the Turkic languages.

The next problem is the development of basic NLP tasks for each language, such as stemming, segmentation, morphological analysis based on the development of specific CSE-model of morphology.



Contribution of the work

The scientific contribution of this work is the development of a CSE-model of the Turkish language and conducting stemming experiments using universal programs of the CSE model approach.

Inferring Of Turkish Complete Set Of Endings

In first step, we have to inferring of acceptable placements of Turkish suffixes types.

The system of endings to the nominal bases of words of the Turkish language has four types:

- plural endings (denoted by K)
- possessive endings (denoted by T)
- case endings (denoted by C)
- personal endings (denoted by J)
- the stem is denoted by S.

Inferring Of Turkish Complete Set Of Endings

Let's consider all possible options for placing types of endings: from one type, from two types, from three types and from four types. The number of placements is determined by the formula:

$$A_n^{\,k} = n!/(n-k)!$$

Then, the number of placements will be determined as follows:

$A_4^1 = 4!/(4-1)! = 4$,

$A_4^2 = 4!/(4-2)! = 12$,

$A_4^3 = 4!/(4-3)! = 24$,

$A_4^4 = 4!/(4-4)! = 24$.

There are **64** possible placements.

Two types

Placements of two types of endings can be as follows:

KT, **TC**, **CJ**, JK
KC, **TJ**, CT, JT
KJ, TK, CK, JC.

the number of valid (correct) placements from two types of endings will be **6**.

KT, TC, CJ, KC, TJ, KJ.

Example:

Araba + lar + ım = Arabalarım (my cars)

Anne + ler + im = Annelerim (our mothers)

Three types

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.

there will be **4** admissible placements of endings of three types .

KTC, KTJ, TCJ, KCJ

Goz+ lar + ım+ da = Gozlarımda(in my eyes)
Anne + ler + in + iz = Anneleriniz (your mothers)

Four types

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

KTJC, TKJC, CKTJ, JKTC

KTCJ, TKCJ, CKJT, JKCT

KJTC, TJKC, CTKJ, JTKC

KJCT, TJCK, CTJK, JTCK

KCTJ, TCJK, CJKT, JCKT

KCJT, TCKJ, CJTK, JCTK

Then, the admissible placements of the four types of endings will be **1**.

KTCJ

Araba+ lar + ım + da + sın = Arabalarımdasın (you are in my cars)

15 admissible placements

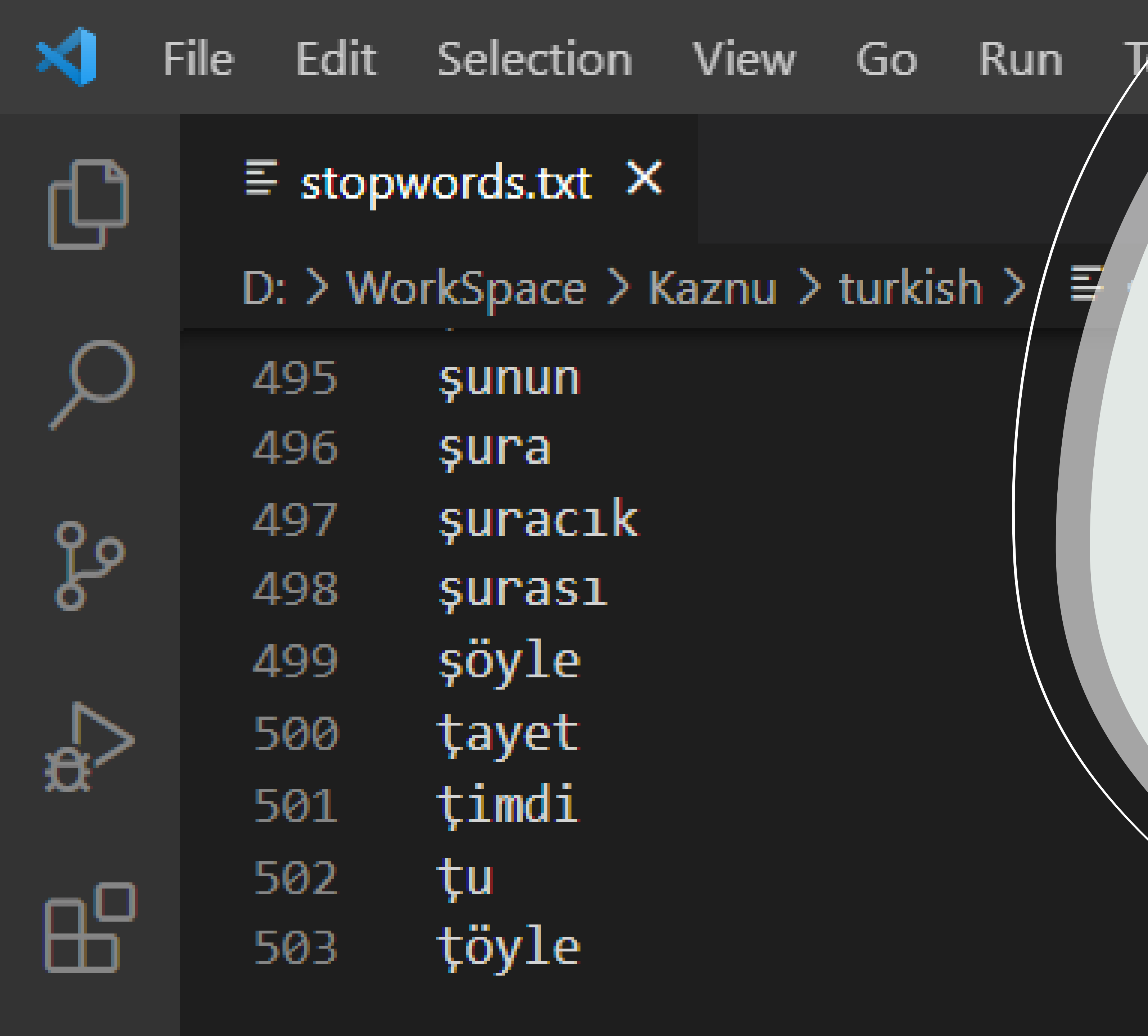
there are 4 admissible placements from one type, 6 from two types, 4 from three types, 1 from four types.

So, the total number of types of allowed placements in words with nominal stems is 15.

Complete set of Turkish endings

According to the above method, we got **3246** complete set of Turkish endings.

Endings	Number
Nominal base	1247
Verbs	427
participles	1299
Adverbs	27
Moods	82
Voices	132
Other Endings	33



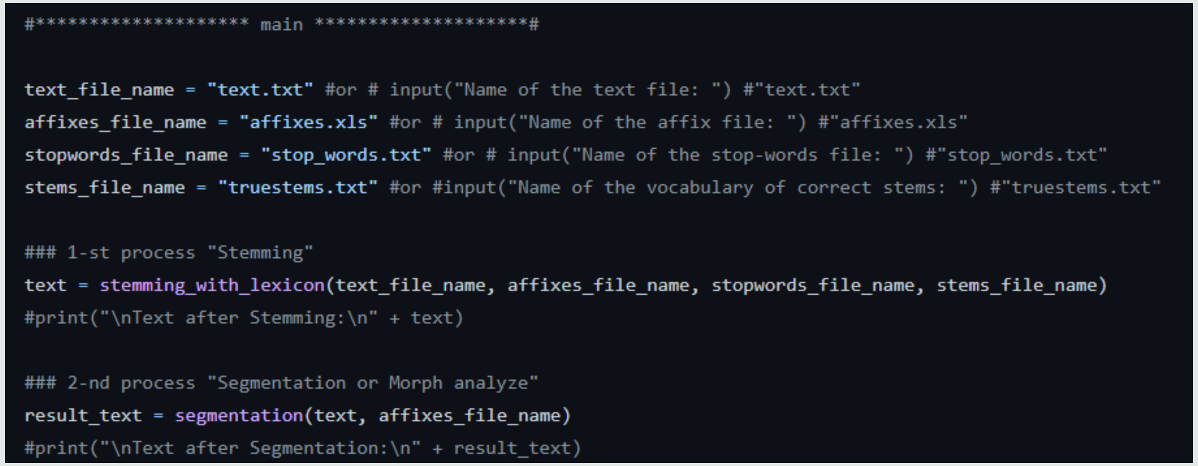
```
495    şunun
496    şura
497    şuracık
498    şurası
499    şöyle
500    ţayet
501    ţimdi
502    ţu
503    ţöyle
```

Stop words

We have also created a **503** word stop-words in networks and books

Stemming

During the research, a program was developed a stemming system in the Turkish language. In the Google Colab application, **35,317** stem words in the Turkish language was compiled, using code written in the Python programming language.



```
#******************** main ********************#

text_file_name = "text.txt" #or # input("Name of the text file: ") #"text.txt"
affixes_file_name = "affixes.xls" #or # input("Name of the affix file: ") #"affixes.xls"
stopwords_file_name = "stop_words.txt" #or # input("Name of the stop-words file: ") #"stop_words.txt"
stems_file_name = "truestems.txt" #or #input("Name of the vocabulary of correct stems: ") #"truestems.txt"

### 1-st process "Stemming"
text = stemming_with_lexicon(text_file_name, affixes_file_name, stopwords_file_name, stems_file_name)
#print("\nText after Stemming:\n" + text)

### 2-nd process "Segmentation or Morph analyze"
result_text = segmentation(text, affixes_file_name)
#print("\nText after Segmentation:\n" + result_text)
```

Experiments and Results

The accuracy of the model in the Turkish language was **87%**.

Resource	Number of Words	Rights	Wrong
Text1	420	397 (87%)	23 (13%)
Text2	626	598 (86%)	28 (14%)

Conclusions

1. New Turkish morphology model on CSE-model
2. Stemming results
3. Future using for preprocessing stage of Neural machine translation

THANK YOU
FOR YOUR
ATTENTION