2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE) | 979-8-3315-1632-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/PIERE62470.2024.10805062

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

Ualsher Tukeyev
Information systems department
Al-Farabi Kazakh National University
Almaty, Kazakhstan
ualsher.tukeyev@gmail.com

Nargiza Gabdullina
Information systems department
Al-Farabi Kazakh National University
Almaty, Kazakhstan
gabdullinanargiza7@gmail.com

Nazerke Karipbayeva
Information systems department
Al-Farabi Kazakh National University
Almaty, Kazakhstan
karipbayeva.nazerke@gmail.com

Nilufar Abdurakhmonova
Department of Computational and Applied Linguistics
National University of Uzbekistan
Tashkent, Uzbekistan
n.abduraxmonova@nuu.uz

Tolganay Balabekova
Information systems department
Al-Farabi Kazakh National University
Almaty, Kazakhstan
t.balabekova@mail.ru

Aidana Karibayeva
Information systems department
Al-Farabi Kazakh National University
Almaty, Kazakhstan
a.s.karibayeva@gmail.com

***Abstract*—The Uzbek language belongs to the Turkic-speaking group and is one of the low-resource languages. In this regard, increasing and expanding the language and electronic resources in the Uzbek language is essential. For many natural language processing (NLP) tasks, such as stemming, segmentation, and morphological analysis, a set of endings and stem and stop words are required. The article contains a complete set of Uzbek endings and a dictionary of stem and stop words. The endings were collected 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. Using the collected linguistic resources, stemming programs for Uzbek texts were tested, problems were identified based on the experiment results, and the program was processed according to them. The results of the experiments using the developed linguistic resources of the Uzbek language showed an accuracy of 94.18% on average.**

Keywords*—computational, model, morphology, Uzbek language, stemming

I. Introduction

Software tools that automatically find the necessary word forms in the studied texts provide substantial support in linguistic research. This problem is partly solved by special programs that search for phrases using the linguistic markup of the corpus texts made in advance. An important, almost central-forming link in the chain of automatic text processing in natural language is the technology of finding the basis of a word (stemming).

Morphological analysis of languages based on two-level morphology uses the apparatus of the theory of automata, namely the apparatus of the theory and methodology of finite transformers (FST). When applying the FST technology for inflected languages, it is essential to determine the set of possible word endings of the analyzed languages. The set of possible word endings will be represented by rows in the multivalued mapping table. If the set of possible word endings of the analyzed language is incomplete, then such a word will be interpreted incorrectly or unsuccessfully. In connection with the above, the problem of determining the complete set of possible word endings of the analyzed language is significant.

This paper represents the computational morphology model and word stemming from the Uzbek language based on a complete set of endings (CSE) morphology models [1].

II. Related Works

The defining approach of morphological analysis is the two-level morphology proposed [2], implemented using finite transducers. Systems of morphological analysis of languages based on two-level morphology use the apparatus of the automata theory, namely, the apparatus of the theory and methodology of finite state transducers (FST). There are publications on two-level morphology technology and FST for agglutinative languages [1]-[9].

The article [3] describes using a two-level morphology apparatus for Turkish. The description of Turkish word morphology uses the PC-KIMMO environment. 22 two-level rules describe the phonetics of modern Turkish. Other exceptional cases of phonological and morphological rules are also considered using a two-level morphology apparatus.

The paper [4] proposes an algorithm for morphological text analysis of Uzbek. The research object of this article is to determine the roots of Uzbek words using morphological rules without additional dictionaries. This article presents an approach to morphological analysis Uzbek text based on word analysis using the finite state machine method to define word roots.

The paper [5] presents a stemming algorithm for the Uzbek language based on an affix-stripping approach. The proposed method of stemming is based on the classification of word affixes in fifteen groups, and an algorithm using Uzbek morphological rules is developed.

The paper [6] proposes a method for analyzing Turkish words without relying on a dictionary. This method uses a rule-based approach and finite state machines (FSMs). Unlike

979-8-3315-1632-1/24/$31.00 ©2024 IEEE

previous studies, the FSMs are created in reverse order based on morphological rules. The paper outlines the steps involved in this process, including categorizing suffixes and generating FSMs for each category.

The article [7] describes the morphemic structure of Uzbek nouns. Using finite automation, an algorithm for morphemic parsing is developed.

The paper [8] presents using the finite-state description to the Kazakh nominals. It develops and implements a finite-state transducer for the nominals of the Kazakh language. It considers the morphophonemic restrictions imposed by the harmony of the Kazakh language on letter combinations in affix connection. The developed Kazakh final transformer implements some functions of morphological analysis.

The article [9] presents a detailed computational analysis of the Kazakh language. The presence of a morphological analyzer is a significant problem, especially for tasks related to NLP in agglutinative languages. The morphological analyzer uses a two-level morphology approach with Xerox finite state tools.

The article [10] developed a set of Uzbek endings based on three inflectional types of nominal base words without personal affixes and two types of derivational affixes. In this paper, the authors present the set of Uzbek endings on the complete set of endings for the different NLP tasks.

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

A. *Inferring of Endings for Nominal Base Words*

The Uzbek language's nominal bases words has four types of suffixes:

- plural suffixes (K),

- possessive suffixes (T),

- case suffixes (C),

- personal suffixes (J),

- the stem is denoted by S.

The number of different placements of suffix types is defined as: for one type 4; for two type 12; for three type 24; for four types 24 [1]. In common, there are 64 possible placements.

Let's consider semantically valid possible placements of suffix types.

Placements of one type of suffixes (K, T, C, J) are all semantically valid by definition.

Placements of two types of suffixes can present as:

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

Placements semantical valid are KT, TC, CJ, KC, TJ, and KJ, and the rest are classified as invalid.

Placements of the three types of suffixes:

KTC, KTJ, KCJ, KCT, KJT, KJC, TCJ, TCK, TJK, TJC, TKC, TKJ, CJK, CJT, CTK, CTJ, CKT, CKJ, JKT, JKC, JTK, JTC, JCK, JCT.

The admissible placements of three types of suffixes are defined according to the rule:

if a placement contains an invalid placement of two suffix types, then that placement is not valid.

The admissible placements of the three suffixes are KTC, KTJ, KCJ, and TCJ.

All placements for four suffix types are:

KTJC, KTCJ, KJTC, KJCT, KCTJ, KCJT, TKJC, TKCJ, TJKC, TJCK, TCJK, TCKJ, CTKJ, CKTJ, CKJT, CTJK, CJKT, CJTK, JKTC, JKCT, JTKC, JTCK, JCKT, JCTK.

The admissible placements of four types of suffixes are defined according to the rule:

if a placement contains an invalid placement of three suffix types, then that placement is invalid.

The admissible placement of the four types of suffixes is KTCJ.

Enumeration of the numbers of Uzbek endings below will infer all valid types of endings. Table I presents the number of endings for the placements of one type of suffix.

TABLE I. Endings of One Type Suffixes.

Suffix type	Suffixes	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

Enumeration of the numbers of Uzbek endings for placement type KT.

Enumeration of placements KT: K * T = 6 (only one plural suffix * possessive suffixes starting with a vowel). Enumeration of the numbers of Uzbek endings for KT is presented in Table II.

TABLE II. Enumeration of the KT Placements.

Example	suffix type K	suffixes type T	Number of endings
aka-	-lar	-im, -ing, -i, -imiz, -ingiz, -niki	6

Enumeration of the numbers of Uzbek endings for placement type KC.

Enumeration in placements KC: K * C = 5. Enumeration of the numbers of Uzbek endings for KC suffix placements is presented in Table III.

TABLE III. ENUMERATION OF THE KC PLACEMENTS.

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

Enumeration of the numbers of Uzbek endings for KJ.

Combinations in placements KJ: K * J = 3. Enumeration of the numbers of Uzbek endings of the KJ suffixes placements is presented in Table IV.

TABLE IV. ENUMERATION OF THE KJ PLACEMENTS.

Example	Suffix type K	Suffixes type J	Number of endings
ona-	-lar	-miz, -siz, -dir	3

Enumerate the numbers of Uzbek endings for type TC.

Combinations in placements TC: T * C = 55. Enumeration of the numbers of Uzbek endings of the TC suffixes placements is presented in Table V.

TABLE V. ENUMERATION OF THE TC PLACEMENTS.

Examples	Suffixes type T		Suffixes type C	Number of endings
	Last sound			
	vowels	consonants		
	-m, -ng, -si, -miz, -ngiz, -niki	-im, -ing, -i, -imiz, -ingiz	-ning -ga -ni -da -dan	11*5=55
ona-	-m, -ng, -si, -miz, -ngiz, -niki		-ning, -ga, -ni, -da, -dan	6*5=30
kitob-		-im, -ing, -i, -imiz, -ingiz	-ning, -ga, -ni, -da, -dan	5*5=25

Here, we look at the word's stem. The endings are selected depending on the last sound in the stem.

Combinations in placements TJ: T * J = 46. There are three types of T*J combinations:
T1-J2:8; T2-J1:8; T1-J3:8; T2-J3:4; T3-J1:4; T3-J2: 4; T3-J3:4; T4-J1, J2, J3: 6.

Enumeration of the numbers of Uzbek endings of placement KTC: 1 type of plural, six types of possessives, and five types of case suffixes are used. The enumeration of the number of Uzbek endings of KTC is 30 (1*6*5=30).

The endings of type placement KTJ have 1 type of plural ending and eight combinations of possessive and personal endings: T1-J2: 4, T1-J3: 2, T2-J1: 4, T2-J3: 2, T3-J1: 2, T3-J2: 2, T3-J3: 1, T4-J1, J2, J3: 5. The possible number of endings of KTJ is 22.

Combinations in placements KCJ: K*C*J = 15. Enumeration of the numbers of Uzbek endings of the KCJ placements includes 1 type of plural, three types of case endings, and five types of personal endings.

The enumeration of the numbers of Uzbek endings of placement TCJ comprises three affixes of type T depending on the word stem, three of C, and four of J. The number of endings of TCJ is 72.

So, there are 339 endings for the nominal base words of Uzbek.

The Uzbek verbs includes three tenses: future, present, and past. Present Tense - hozirgi zamon davom fe'li (in Uzbek language). The Present tense verb is formed by three different endings with the root. The first way is attaching the affix -yap to the word stem. The affix forms the second way **-**moqda. The third way uses affixes -ayotir and -(y)yotir. The third way of forming a present continuous tense verb is shown in Table VI.

TABLE VI. THE THIRD METHOD OF FORMING THE PRESENT CONTINUOUS TENSE VERB.

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

The ending negative form – ma + endings of the present continuous verb. The present continuous tense has 76 possible endings, including negative affixes and question forms.

The present–future tense is formed using the endings -a (after constant) and -y (after a vowel). For example: men kel-a-man (I will come). The ending forms the negative form ma + y (ending of the present-future verb). This tense has 42 endings in question form. The number of present tense endings is 118.

There are four types of past tense. First is O'tgan zamon aniq fe'li (in Uzbek language). A simple past tense verb combines the ending -di with the root. For example: men kel-di-m (I came). This tense has 28 endings with negative affixes and question affixes. The second type is Yaqin, a verb that combines the ending -gan with the root. A negative form is formed by attaching the affix -ma. For example: uxla-ma-gan-san (you haven't slept). The second negative form is verb + -gan emas+(personal suffixes); therefore, this type is not considered. Examples: kelgan emasman, kelgan emassan. This type of past simple has 28 endings. The third type is

O’tgan zamon davom fe’llari (in Uzbek language). This tense is formed with two methods: 1. It is formed using the suffix -(a)r and the verb edi. Examples: men kelar edim, men kelmas edim (personal suffixes). 2. It is formed using the suffix -(a)yotgan + -di + personal suffix. An example of the second formation of a past tense verb is presented in Table VII. The third type of past simple has 21 endings with its negative form.

TABLE VII. EXAMPLE OF THE FORMATION OF THE PAST TENSE VERB.

Examples	Suffixes	1st person	2nd person	2nd person (respect)	3rd person	Number of endings
kel- yoz-	-ayotgan+ -di	-m	-ng	-ngiz	-	4
uxla- o’qi-	-yotgan + -di	-k	-ngiz	-ngizlar	-lar	3 7*2=14

The fourth type of past simple is O’tgan zamon hikoya fe’llari. A verb of this type of past tense: the root + the ending -(i)b + personal suffix. For example, men kel-ib-man (I came). This type has 45 endings with negative and question affixes. The number of past tense endings is 122.

The future tense in the Uzbek language has two types. There are future-specific tenses and future-possible tenses. Kelasi zamon maqsad fe’llari (in Uzbek language)- future specific tense. A future-specific tense verb combines the ending -(a)r + personal suffix with the root. For example: men kel-ar-man (I will come). The formation of a verb of of future specific tense is presented in Table VIII.

TABLE VIII. EXAMPLE OF THE FORMATION OF THE FUTURE SPECIFIC TENSE VERB.

Examples	Suffixes	1st person	2nd person	2nd person (respect)	3rd person	Number of endings
kel- yaz- kut-	-ar	-man	-san	-siz	-dir	4
so'ra-	-r	-miz	-siz	-sizlar	-lar	2 6*2=12

The ending forms the negative form -mas + personal suffix. For example: men kel-mas-man. This type os future tense has 33 endings with its question form. Kelasi zamon gumon fe’llari (in Uzbek language) - future possible tense. A verb of future possible tense is formed by combining the ending -moqchi + personal suffix with the root. For example, men kel-mochi-man (I want to come). Negative form – verb root + moqchi emas + personal suffix. This type has 20 endings. So, the number of future tense endings is 53.

So, the number of verb endings in Uzbek language is 293.

The participle's endings begin with base suffixes (R). included next 12 suffixes: -gan, -qan, -kan, -ayotgan, -yotgan, -adigan, -ydigan, -ar, -r, -mas, -ajak, -mish.

Considering the possible sequences of suffix types for participles (the base affixes are the same for all variants), the following sequences are semantically permissible:

- One type affix: RK, RT, RC, RJ
- Two type affixes: RKT, RKC, RKJ, RTC, RCJ,
- Three type affixes: RKTC, RKCJ

Therefore, there are 12 permissible types of participle endings.

The affixation of the word “chiq” (get out) with RK endings will be as follows: chiq-qan-lar (those who came out). The total number of inferring endings for participle placements type RK is 12.

RKT has 12 forms of base suffixes. Table IX shows the enumeration of RKT endings.

TABLE IX. ENUMERATION OF ENDINGS FOR RKT.

Examples	Suffix R	Suffix K	Suffixes T	Total number of endings
yoz chiq o`t yoz uxla yoz uxla yoz so'ra kut kel ket	gan qan kan -ayotgan yotgan adigan ydigan ar r mas ajak mish	-lar	-im -ing -i -imiz -ngiz	12*5=60

Thus, the total number of participle endings in the Uzbek language is 1344.

In the Uzbek language, verbs have three moods: indicative, imperative, and conditional. The endings of the indicative mood correspond with the endings of the tense verbs, so we do not consider them. Imperative mood is verbal stem + the affixes -ay, -aylik, -y, -ylik, -gin, -ing, -ingiz, -ingizlar, -ng, -ngiz, -ngizlar, -sin, -sinlar. The total number of imperative moods is 13. Examples of the formation of the imperative mood are given in Table X.

TABLE X. THE FORMATION OF THE IMPERATIVE MOOD OF THE VERB.

Examples	1st person	2nd person	3rd person	Number of endings
yoz-	-ay -aylik	-gin -ing, -ingiz, -ingizlar	-sin -sinlar	13
o’qi-	-y -ylik	-gin -ng, -ngiz, -ngizlar		

The conditional mood is the verbal stem + the endings + the personal suffix. There are 27 mood endings with their negative form.

There are four types of voice endings in the Uzbek language: passive voice, reflexive voice, reciprocal voice, and causative voice suffixes. Passive voice is the verbal stem plus the affixes -l and -il with the verb suffix -di + personal suffix—for example, tara-l-di-m. The total number of voice endings is 252.

Thus, the complete set of endings for nominal-based stems is 339, and for verbal stems are: verb– 293, participles – 1 344, moods – 27, and voices – 252. In total, there are 2255 endings in the Uzbek language.

IV. Experiments and Results

The section presents the results of Uzbek stemming experiments by the universal program for the model [11]. We created a vocabulary of the Uzbek language consisting of 590 stop words and 23000 stems. As a resource, we took from the Internet texts about computers and books in the Uzbek language. We then experimented with using these accumulated linguistic resources and a stemming algorithm with the stem lexicon. In total, 82 sentences comprising 1,046 words were experimented. We manually checked each word of the result obtained separately with the help of the experiment. The model's accuracy with the texts of the Uzbek language is presented in Table XI.

TABLE XI. Percentage Analysis of the Stemming Result

Resource	Words count	Correct count	Incorrect count
1-text	420	394 (93,8%)	26 (6,2%)
2-text	626	592 (94,57%)	34 (5,43%)

V. Conclusions and Future Works

The article presents several new resources on the Uzbek language: the set of endings and the dictionary of stem and stop words. The endings were collected 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. The result of the experiment using the accumulated linguistic resources showed an accuracy of 94.5%. In the future, the dictionary of endings and stems of the Uzbek language will be used in the morphological analysis of Uzbek texts, text segmentation, machine translation, and generation of datasets in neural machine translation.

Acknowledgments

This research was performed and financed by the grant Project IRN AP23487816 of the Ministry of Science and Higher Education of the Republic of Kazakhstan.

References

[1] U. Tukeyev, Automaton models of the morphology analysis and the completeness of the endings of the Kazakh language. Proceedings of the international conference “Turkic languages processing”, 91-100 pp. TURKLANG-2015 September 17–19, Kazan, Tatarstan, Russia, 2015.

[2] K. Koskenniemi, Two-level morphology: A general computational model of word-form recognition and production. Tech. rep. Publication No. 11. Department of General Linguistics. University of Helsinki, 1983.

[3] K. Oflazer, Two-level description of Turkish morphology, Literary and Linguistic Computing Volume9, Issue2. 137-148, 1994.

[4] J.A. Atadjanov, Models of Morphological Analysis of Uzbek Words. Cybernetics and programming № 6.70 – 73pp. DOI 10.7256/2306-4196.2016.6.20945, 2017.

[5] M. Sharipov, O. Yuldashov, UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language. Proceedings of the International Conference on Agglutinative Language Technologies as a Challenge of Natural Language Processing (ALTNLP), June 6, 2022, Koper, Slovenia, CEUR Workshop Proceedings (CEUR-WS.org/vol-3315/paper15), 2022.

[6] G. Eryiğit, E. Adali, An affix stripping morphological analyzer for Turkish. In M. H. Hamza (Ed.), Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (as part of the 22nd IASTED International Multi-Conference on Applied Informatics, pp. 299-30420, 2004.

[7] I. I. Bakaev, T. Shafiyev, Morphemic analysis of Uzbek nouns with Finite State Techniques Journal of Physics Conference Series 1546:012076, May 2020.

[8] B. M. Kairakbay, D. L. Zaurbekov, Finite State Approach to the Kazakh Nominal Paradigm. Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing. 108–112 pp. St Andrews–Sctotland, 2013.

[9] G. Kessikbayeva, I. Cicekli, Rule Based Morphological Analyzer of Kazakh Language. Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM, Baltimore, Maryland USA . 46–54 pp., 2014.

[10] S. Matlatipov U. Tukeyev, M. Aripov, Towards the Uzbek Language Endings as a Language Resource. In: Hernes M., Wojtkiewicz K., Szczerbicki E. (eds) Advances in Computational Collective Intelligence. ICCCI 2020. Communications in Computer and Information Science, vol 1287, pp.729-740. Springer, Cham. DOI 10.1007/978-3-030-63119-2_59, 2020.

[11] NLP-KAZNU: github.com/NLP-KAZNU