Morphemic analysis of Uzbek nouns with Finite State Techniques

I I Bakaev¹, T R Shafiev²

¹Tashkent University of Information Technologies, 108, Amir Temur ave., Tashkent, 100200, Uzbekistan
²Bukhara State University, 11, M Ikbol Str., Bukhara, 200114, Uzbekistan
E-mail: bakayev2101@gmail.com

Abstract. The article developed a mathematical model that describes the morphemic description of the structure of nouns in the Uzbek language. The developed model takes into account all forms of word-forming suffixes, affixes and creating word forms that belong to the lexical category of nouns. For the created mathematical model of word forms, an algorithm for morphemic parsing and determining the pre-ownership of word forms using finite automation is developed.

1. Introduction

The huge flow of information on the internet has led to the rapid development of the natural language processing (NLP) industry. Various research engines are currently developing their projects, such as information exchange between users, machine translation of information, checking e-mails for spam filters, and processing question-and-answer systems. However, there are problems that due to insufficient knowledge of the structure of some languages, the research result does not fully meet the user's needs.

Today, one of the problems of research engines is the morphological and morphemic analysis of words that is encountered when processing user requests. An example of such languages is Uzbek, which belongs to the family of Turkish languages. Due to the influence of Arabic, Persian, and Russian, the structure of the language has become more complex. To solve such problems, we use morphological analysis and morphemic analysis of words, which gives more accurate results on search queries.

In the field of computer linguistics, tasks for morphological analysis are occupied by scientists such as Ermakov A. E, Zygmunt Vetulani, V. B. Barakhnin, A. Ismailov, Zailani Abdullah, Masita Abdul Jalil, Noor Hafizah Abdurahim, Koibagarov G. Ch., Musabaev R. R., Kalimoldaev M. N., Fedotov A. M, Tusupov D. A, Sambetalbayeva M. A, Erimbetova A. S, Bakieva A. M, Idrisova I. A, and others. also in our country, this area is actively studied by our own scientists Matlatipov G, Abdurakhmonova N, Tuliyev U, Atadjanov J. A. and currently these researchers had obtained some results.

In particular, the author [1] considers the linguistic properties of the Uzbek language and developed a text analysis algorithm. Also, the article discusses the categories of suffixes in the Uzbek language and their rules for the use of suffixes in texts. Using a finite state machine and graphs [2], parsing is described and formal formulas are proposed for determining the roots of words. This algorithm determines plagiarism. And the method of mathematical modeling is also considered.
In the article [3] grammatical categories of verbs of the Uzbek language are studied. When creating a grammatical category, segmentation is used, this is a multi-level process for morphological analysis. The authors conducted research to create grammatical and spelling rules that are important for the final state of the Converter, and the article presents some ways to solve the problem of morphological-analyzer for machine translation.

The author’s article [4] considers the significant role of morphemic analysis for modeling grammatical categories of speech parts in the Uzbek language in machine translation. It is determined to identify the types of stages of morphological analysis and common paradigms, differences between the source and target languages. The author shows that verbs in the Uzbek language have their own characteristics in terms of forms and analytical characteristics, as well as a complex verb, a phrase; verbal phrases and their forms play an important role in computational morphology. As a result of the study, the modeling of grammatical categories based on forms, restrictions of syntactic settings and combinations of affixes in verb forms is shown.

In [5], we created an algorithm for machine translation from English into Uzbek and from Uzbek into English. For machine translation, at the first stage, the method of morphological analysis was used, such as tokenization (parsing by words), lemmatization (morpheme analysis), and stemming (morphological search highlighting the word base allows the search engine to search for a word not in a strictly specified form, but in all its morphological forms). In the second stage, the syntactic models of the text were compared with each other.

The article [6] describes the morphological analysis of the Uzbek language using the logical programming language Prolog. The authors of the article implemented a parsing tool that performs segmentation of Uzbek words into roots and suffixes. The developed algorithms for word segmentation relate to the written Uzbek language, which corresponds to the spelling principles set out in the document "Basic spelling rules" for the Uzbek language. Resolution Of The Cabinet Of Ministers Of Uzbekistan No. 339. The developed algorithm created the system of automatic morphological analysis of the Uzbek language UZMORPP.

In this research [7] compares existing stemmers, and their use for the Uzbek language. The authors discuss the type of stemming algorithms, including a review and comparison of popular English stemming, evaluated and analyzed the available stemmers for predicting experiments in the Uzbek language. Based on a comparative study and experiment, the authors propose an Uzbek stemmer (the program mechanism of morphological search) model that extends some of the functions of the logins stemmer in accordance with the requirements for the Uzbek language.

The article [8] proposes algorithms for the analysis and synthesis of word forms of the Kazakh language, which also refers to agglutinative languages. These algorithms are based on the principles of word separation. Created dictionaries, including about 2000 affixes and their combinations. The dictionary makes it possible to use the algorithm for 17 classes and 3,500 affixes for the lexical categories of nouns and adjectives. These algorithms are used for morphological analysis in search engines, auto-references and in the construction of thesauruses and ontologies.

The authors of [9] analyzed the algorithms to normalize the words of the Kazakh language. The developed algorithms have the ability to automatically select the basics and methods for the synthesis of normal word forms for the Kazakh language are described. Algorithms are implemented by processing word forms using a dictionary and without a dictionary. A thesaurus of scientific and technical terms in the field of information technology in the language of punctuation has been created. This proves the effectiveness of the developed algorithm.

For morphological analysis, the authors [10] proposed a declarative, procedural, and hybrid approach. In the declarative approach, morphological information is stored in two dictionaries, such as the dictionary of lemmas and affixes. It is shown that, in the procedural approach, each word is divided into a word base (root), a chain of affix bases and organized as a dictionary. The above approaches are implemented on the basis of the theory of finite automata. The authors also describe the principle of operation of the morphological analyzer using the schematic method.
A specific way to solve the problem of finding the basis of words is called a stemming algorithm, and a specific implementation is called a stemmer [11].

Based on the above, the purpose of this article is a detailed analysis and development of an algorithm for morphological and morphemic analysis of the Uzbek language, which takes into account the word formation of affixes (suffixes), affixes, affixes and other forms of suffixes for nouns. As well as the development of a qualitative algorithm for the description of morphemic parsing of nouns and morphological analysis of the Uzbek language from finite automata.

2. Specification of Uzbek nouns

The Uzbek language is one of the agglutinative languages of the Turkic family. In this language, each grammatical meaning is expressed in separate affixes [12]. The term affix in the grammars of the Uzbek language, as in the grammars of other Turkic languages, is taken in a generalized sense. It means prefixes, infixes, suffixes, endings, prefixes [13]. In Uzbek, word-forming affixes (suffixes) are usually added after the root of words, and then endings are added to words (affixes of belonging, case affixes).

Rules for the relationship of the Uzbek noun and affixes

Let's introduce some conditional variables to show the rule of relationship between the Uzbek noun and affixes

\[ W \rightarrow \text{word form, } W = \{w_1, w_2, \ldots, w_n\} \]

\[ K \rightarrow \text{root of the word} \]

\[ SX \rightarrow \text{word-forming affixes (suffix)} \]

\[ AF \rightarrow \text{affixes (subjective markdown affixes, plural affixes, accessories, case affixes)} \]

For example,

- vatan+dosh (compatriot) root+ {suffix}
- do'st+lik (friendship) root + {suffix}
- kitob+lar (books), root+ {affix}
- daftar+im (my notebook), root + {affix}

Let \( + \) - operation concatenation of character strings

\[ \sum \rightarrow 0, 1, \ldots n \text{ consecutive operation concatenation.} \]

Formally describing the mathematical model of education for word forms of nouns:

\[ W_1 = K + \sum_{j=0}^{4} s x_j + \sum_{i=0}^{n} a f_i \]  \hspace{1cm} (1)

vatan+dosh (compatriot) root+ {suffix}

do'st+lik (friendship) root + {suffix}

\[ W_2 = K + \sum_{i=0}^{4} s x_j \]  \hspace{1cm} (2)

kitob+lar (books), root+ {affix}

daftar+im (my notebook), root + {affix}

\[ W_3 = K + \sum_{i=0}^{n} a f_i \]  \hspace{1cm} (3)

In the Uzbek language, suffix-word-forming morphemes are attached in a certain sequence. Sometimes the number of morphemes in a word form can be more than two to three or four [14]. For example,

- til+ shunos+lik,
til - the language of,

til+shunos- linguist

tilshunos+lik- language studies

root+ {suffix}+{suffix}

In the above examples, the "right-hand rule" is applied, that means the affixes are added from the right [12].

3. Method of solution

Generically, you can describe the formation of nouns using finite state machine (FSM) in the following form:

$$M = (W_n, Q, q_0, F, \delta),$$

finite state machine: $M$

set of Input alphabet: $W = \{K, SX, AF\}$

Finite set of states: $Q = \{q_0, q_1, q_3\}$

Transition Function: $\delta(Q, W): Q \times W \rightarrow Q$

Initial state: $q_0$ , ($q_0 \in Q$);

set of Final States: $F$ , ($F \subseteq Q$);

![Figure 1. FSM for parsing nouns](image)

We describe the value of the transition function for FSM nouns:

$$\delta(q_0, K) = q_1$$

$$\delta(q_1, SX) = q_2$$

$$\delta(q_1, AF) = q_3$$

$$\delta(q_2, AF) = q_3$$

$$\delta(q_3, AF) = q_3$$

As can be seen from the above example, it allows word forms of nouns with the formula (1). FSM allows word forms when the state is $q_2, q_3$.

You can convert this FSM to a FSM with a single end state:

![Figure 2. This example is for parsing nouns with a single FSM](image)

Description of morphemic selection of nouns using finite automata
The word form $W$ is allowed using FSM if

$$ W = W_1, W_2, \ldots, W_n $$

If $\exists r_0, r_1, \ldots, r_n$: sequences of States that

$$ r_0 = q_0 $$
$$ r_i = \delta(r_{i-1}, W_i), i = 1 \ldots n $$
$$ r_n \in F $$

The value of the transition function for the given FSM:

$$ q_0 \text{ - Initial state} $$
$$ \delta(q_0, K) = q_1 $$
$$ \delta(q_1, SX) = q_2 $$
$$ \delta(q_1, \varepsilon) = q_2 $$
$$ \delta(q_2, SX) = q_2 $$
$$ \delta(q_2, AF) = q_3 $$
$$ \delta(q_2, \varepsilon) = q_3 $$
$$ \delta(q_3, AF) = q_3 $$

### Table 1. State-transition table:

| State of the FSM | K | SX | AF |
|------------------|---|----|----|
| q₀               | q₁ | -  | -  |
| q₁               | -  | q₂ | -  |
| q₂               | -  | q₂ | q₃ |
| q₃               | -  | -  | q₃ |

In the Uzbek language, according to speech requirements, words can take different suffixes, while the meaning of the word changes and expresses different meanings. This is especially true for words in the noun category.

For example, let's analyze the words ishshi, ishxona, ishlar, ishi, ishdan. In these examples, you can highlight the root of ish (work). This for all the above word forms is considered the basis. – chi, -xona, -lar, -i, -dan-affixes.

chi, xona-word-forming affixes
lar is the plural affix
I-affix of affiliation
dan-case affix

As a result, the word form has the following structure and washed away:

Ish+chi (worker) (noun + suffix= name of the actor)
Structure (S) = root+ {suffix}

ish+xona (workplace) (noun + suffix = locality)
S = root+ {suffix}

In the word formation of nouns, not only word-forming affixes are involved, but also affixoids. In this example, xona is considered an affixoid.

Affixoid is a common name for special morphemes of the transitional type that occupy an intermediate position between roots and affixes [15].

ishi+lar (cases) (noun + plural= plural noun)
S = root+{affix}
ish+i(his (her) Affairs)(noun + affix of affiliation = noun of 3 person singular)
example: uning sizda ishi bor - he has a case for you
S = root+{affix}

ish+dan (from work) (noun + original case= noun in the original case)
S = root+{affix}

FSM, which allows all word forms formed on the basis of the root "ish", has the following form:

Figure 3. FSM for «ishchi, ishxona, ishlar, ishi, ishdan».

In the process of development of society, new concepts appear. These concepts are expressed in separate words. The possibilities of the Uzbek language are so rich that they can Express a variety of information using their affixes. In the Uzbek language, morphemes (suffixes) are used to form new concepts and form new words [12]. Creating new words by adding a suffix always requires two parts:

The basis for creating word forms: (these are independent parts of speech " Lexical category": ot (noun), sifat (adjective), son (numeral), olmosh (pronoun), Fe'l (verb), ravish (adverb). In addition, there are three special word groups in the Uzbek language: undov (interjection), modal so'z (modal words), and taqlidiy so'z (imitative words) serve as the basis for creating a word)

Word-forming part: (these are word-forming affixes). The basis of word formation of nouns is the following:

- Kitob+xon (reader)-[noun+suffix= noun]
- Yaxshi+ lik (dobrata)[adjective + suffix= noun]
- Bir+lik (unity) [numeral name + suffix= noun]
- O'z+lik[noun +suffix= noun](for example, myself) (my me for example,
- Keldi ochilur chog'i, o'zliging namoyon qil.(Xamza.)- It's time for emancipation, show (show) your essence )

As can be seen from the above examples, there is a group of word-forming affixes that creates word forms belonging to the lexical category of nouns in the Uzbek language. Word-forming affixes are divided into four groups.

Word-forming affixes of the actor
Word-forming affixes of nouns with the meaning of a place
Word-forming affixes of nouns with the meaning of the subject
Word-forming affixes of abstract nouns

Group 1: fol+ bin, darvoza+ bon, masxara+ boz, non+ voy (novvoy). o’qi+ vchi, yoz+ uvchi, zar+ gar, fazo+ gir, maslahat+ go’y, chorva+ dor, vatan+ dosh, etik+ do’z, arava+ kash, paxta+ kor, go’r+ kov, tarix+ navi, osh+ paz, kema+ soz, baliq+ furush, kitob+ xon, meros+ xo’r, ish+ chi, til+ shunos, ham+ yurt

Group 2: do’mbir+ a, qars+ ak, ko’k+ at, ko’ch+ at, soya+ bon, surur+ gi, tut+ qich, chiz+ g’ich, orin+ diq, qalam+ don, o’s+ imlik, ek+ in, tug+ un, chiri+ ndi, chiq+ indi, yuv+ undi, chiq+ it, bog’+
ich, ela+ k, to’pla+ m, o’s+ ma, il+ moq, taklif+ nomax, belan+ chak, o’yin+ choq, yopin+ chiq, qalam+ cha.

Group 3: ziyorat+ goh, olcha+ zor, gul+ iston, tosh+ loq, dars+xona

Group 4: och+ archilik, odam+ garchilik, sez+ gi, quv+ g’in, bola+ lik, bor+ liq, o’t+ mish, kelish+ movchilik, she’riy+ at, sevin+ ch, qo’rq+ inch, tushun+ cha, hunarmand+ chilik, arzon+ chilik, qiyana+ q.

The above groups have identical word-forming affixes, which creates word forms belonging to two groups or another lexical category. For example, the affixes “bon”, “cha”, “at”. We describe a word form membership group using a non-deterministic finite state machine (NDFA) in Fig. 4.

Nondeterministic finite state machine (N DFA) in the following form:

\[
M = (W, Q, q_0, F, \delta),
\]

FSM: \( M \)

input alphabet: \( W = \{K, SX, AF\} \), \( K = \{k_1, k_2\} \), where \( k_1=“darvoza”\) («gate»), \( k_2=“soya”\) («shadow»).

Finite set of states: \( Q = \{q_0, q_1, q_3\} \)

Transition Function: \( \delta: Q \times W \rightarrow 2^Q \)

Initial state: \( q_0 \), \( (q_0 \in Q) \);

set of Final States: \( F, (F \subseteq Q) \);

\[ q_0 \xrightarrow{K_1 \text{ bon} \text{ bon}} q_1 \xrightarrow{q_2} q_3 \]

**Figure 4.** NDFA for the description of membership groups of word forms.

The function of transfer NFA:

\( q_0 \) - Initial state

\( \delta(q_0, k_1) = q_1, i = \{1, 2\} \)

\( \delta(\delta(q_0, k_1), “bon” ) = q_2 \)

\( \delta(\delta(q_0, k_2), “bon” ) = q_3 \)

In condition when \( q_2 \) NFA defines word forms “darvozabon” («goalkeeper») which belongs to the 1st group, otherwise, condition \( q_3 \) word form “soyabon” («umbrellas») belongs to the 2nd group. However, when implementing the NFA, it is recommended to use a table of features or a morphological lexicon of word forms.

4. Conclusions

A qualitative and comprehensive analysis of scientific research related to the problem of morphological analysis of the Uzbek language is made. Models and algorithms for using suffixes for agglutinative language families are also studied.

To study the model of morphological analysis of the Uzbek language, a mathematical model is proposed that takes into account word-forming suffixes, affixes, and other forms of affixes for nouns.
Based on the created mathematical model, we developed an algorithm for morphemic parsing and a group of pre-belonging word forms using finite automata, which supports parsing word forms belonging to the lexical category of nouns.

References
[1] Atadzhanov J 2016 Cybern. Program. 6 70
[2] Ashob A and Ulman D 1978 Theory of parsing translation and compilation (Parsing vol 1) (Moscow: Mir) p 612
[3] Abdurakhmonova N and Tuliyev U 2018 Foreign Philol. 3 63
[4] Abdurakhmonova N 2017 Soc. Sci. J. 3 89
[5] Abdurakhmonova N 2017 Machine Translate 1 153
[6] Matlatipov G and Zymunt V 2009 Asp. Nat. Lang. Process. 5070 83
[7] Ismailov A, Abdul Jalil M M, Abdullah Z and Abd Rahim N H 2016 A comparative study of stemming algorithms for use with the Uzbek language Computer and Information Sciences (ICCOINS) (Kuala Lumpur: IEEE) pp 7-12
[8] Baraxnin V B, Bakieva A M, Bakiev M N and Tajibaeva S J 2017 Comput. Technol. 22 11
[9] Fedotov A M, Tusupov D A, Sambetbayeva M A, Erimbetova A S, Bakiyeva A M and Idrisova A I 2015 Bull. NSU. Ser. Inf. Technol. 13 107
[10] Koybagarov K C, Musabayev R R and Kalimoldayev M N 2014 Comput. Sci. 3 64
[11] Anjali G J 2011 Int. J. Comp. Tech. Appl. 2 1930
[12] Nurmonov A, Sobirov A and Qosimova N 2016 Modern uzbek literary language (Tashkent: Ilm Ziyod) p 576
[13] Reshetov V V 1965 The basics of phonetics and grammar of the Uzbek language (Tashkent: Teacher) p 243
[14] Mengliyev B and Bahriddinova B 2007 The vocabulary of the Uzbek language (Tashkent: Yangi asr avlodi) p 590
[15] Gushina L. N 2003 GGMU 2 19