Bulletin of TUIT: Management and Communication Technologies

Volume 2 *Issue 1* Article 2

6-30-2019

MATHEMATICAL MODEL AND ALGORITHM FOR CALCULATING COMPLEX WORDS IN THE KARAKALPAK LANGUAGE

Shaxnoza Abidova
Bulletin of TUIT: Management and Communication Technologies, ab.shaxnoza84@gmail.com

Recommended Citation

Abidova, Shaxnoza (2019) "MATHEMATICAL MODEL AND ALGORITHM FOR CALCULATING COMPLEX WORDS IN THE KARAKALPAK LANGUAGE," *Bulletin of TUIT: Management and Communication Technologies*: Vol. 2 , Article 2.
Available at: https://uzjournals.edu.uz/tuitmct/vol2/iss1/2

UDC 510.5, 519.768.2

MATHEMATICAL MODEL AND ALGORITHM FOR CALCULATING COMPLEX WORDS IN THE KARAKALPAK LANGUAGE

Nazirova E. Sh., Abidova Sh.B.

Abstract. The article examines the morphology of the Karakalpak language, which belongs to the Kipchak group of the Turkic language family. The forms of word formation in the Karakalpak language, their sequences and the affixes added to the core are analyzed. On the basis of the analyzed affixes and suffixes, a complex mathematical model of word formation in the Karakalpak language was developed. On the basis of the developed mathematical model, an algorithm for creating a complex word in the Karakalpak language was developed. Using the developed mathematical model, a four-stage scheme was created for creating complex words of the Karakalpak language.

Keywords: morphology, mathematical model, electronic translator, affixes.

Introduction

As a result of the development of modern information technologies and communications, access to multilingual information is provided. As the flow of multilingual information expands and becomes more complex, information processing, its quality and rapid translation from one language to another remain one of the most pressing issues. It takes a lot of time and money to solve this problem with a dictionary. Therefore, the demand for electronic translators is growing. Especially in the current period of the pandemic, a dictionary or the help of a translator is needed so that people can translate the information, scientific and fiction books they need at home from one language to another. During a pandemic, it is impossible to turn to a translator, and the dictionary is time consuming. Therefore, the use of electronic translators is advisable in all respects.

In translation studies, more and more attention is paid to the analysis of electronic means, which make it possible to speed up and optimize the translation process. Domestic and foreign scientists-linguists, practitioners and theorists-translation studies, especially specialists in the field of translation terminology and machine translation, noting the increasing importance of information technologies in linguistics in general and in translation in particular, are developing various strategies and methods for their most effective application in professional activities [1-6].

The article by A.O. Kazennikov presents a method for removing morphological homonymy. The proposed method combines classical morphological analysis and allows to simultaneously solve the problems of lemmatization and restoration of morphological features [7]. The task of lemmatization is to reduce a word or word form to a lemma or normal form.

Machine translation (MT) is the process of translating written texts from one natural language into another using a special computer program. Sometimes such an appeal turns out to be useful, as it allows you to quickly understand the main content of the text, but much more often, especially when working with scientific and technical literature, such translations do not stand up to criticism [8].

Machine translation can use a method based on linguistic rules, which means that words will be translated in a linguistic way - the most suitable (orally speaking) words of the target language will replace those in the original language.

It is often argued that the success of machine translation requires the problem of understanding natural language to be addressed first.

This article is devoted to morphological analysis, the creation of complex words in the Karakalpak language, a mathematical model and an algorithm for translating complex words.

First of all, the morphological analysis should be briefly described.

Morphological analysis is a procedure as a result of which information about its internal structure is obtained from the form of the external design of a word in the text. Today, there are several dozen morphological analysis algorithms for different languages. The main directions of morphological analysis are:

1. Analysis by dividing a word form into a stem and an intended ending, followed by checking for compatibility.

2. Morphological analysis of the final combination of letters.

3. Universal mathematical models that can be used by the methods of morphology for morphological analysis, in the form of open systems of equations, allowing by calculations to carry out the normalization of word forms, obtaining grammatical information and the synthesis of word forms.

When translating words using electronic translators, morphological analysis of words is performed first. Since words must be separated by root and affixes. To create a machine translator, it is important to know the morphology of the translated languages.

It should be noted that different languages have different semantic and grammatical features, so often algorithms successfully used to process one language show very low efficiency in another language. The complexities of natural language processing, however, do not exclude the possibility of identifying narrower problems that can be solved algorithmically: for example, determining parts of speech or splitting texts into logical groups. However, some features of natural languages significantly reduce the effectiveness of these solutions. Thus, for example, taking into account all word forms for each word in the Karakalpak language increases the complexity of text processing by an order of magnitude. Note that the Karakalpak language, which is part of a large group of Turkic languages, refers to an agglutinative language.

The main feature of languages of the agglutinative type is that the forms of independent words are formed with the help of unambiguous affixes freely attached to the original form. The term ag-glu-tinatio etymologically means "sticking, sticking".

The essential features of agglutinative languages include:

- transparency of the syntagmatic structure of the word, free articulation into morphemes;
- axial (axial) nature of the paradigmatic structure, free constructability of word forms;
- the linear character of the word, the coincidence of the stem with the root and with any word form that serves to build more complex word forms in terms of the number of grammatical meanings.

Agglutinative affixes are characterized by the following features:

- unambiguity: each affix usually expresses one category;
- standard: the affix usually has no variants; free adherence to the word [9-12].

Main part

Review of existing solutions

Consider morphological models, the underlying approaches to the construction of algorithms for normalizing words. The existing approaches are divided into two classes: stemming and lemmatization algorithms.

Stemming is about finding the stem of a word for a given source word. The stem of a word may not always be the same as the root.

Lemmatization is the process of reducing a word (word form) to a lemma (normal form) [13-14].

Morphological analysis provides a solution to two main tasks:

- the problem of analyzing the definition of the normal form of a word by an arbitrary word form;

- problems of synthesis of construction of all word forms in normal form.

Most of the popular algorithms implement lemmatization (reduction to normal form) using the stem of the word, the stemming algorithm [15].

Let's analyze the two most popular lemmatization algorithms based on different principles - Porter and Yandex.

Porter's stemming algorithm was published in 1980 by Martin Porter for the English language [15]. The main idea of Porter's algorithm is that there is a limited number of form and word-forming affixes, and the stem of a word is transformed without using any bases (dictionaries) of stems: only a lot of existing affixes (while complex compound affixes are broken into simple ones) and manually set regulations.

The fact that Porter's algorithm does not use any dictionaries and base bases is a plus for performance and a range of applications (it does a good job with non-existent words), and at the same time a minus in terms of the accuracy of the base selection. In addition, the human factor is often attributed to the disadvantages of Porter's algorithm: the fact that the rules for checking are set manually and are sometimes associated with the grammatical features of the language increases the likelihood of error [13, 15].

The Yandex algorithm (Mystem) is the development of Ilya Segalovich (Yandex, 1998). This morphological analysis algorithm is a dictionary one. The main feature of the algorithm is that for a word form that is not described in the dictionary or a non-existent word, the algorithm generates its hypothetical model of the word change - one or more variants of the normal form of the word then, replenishing the dictionary with new tokens, the generated hypothetical entries can be saved in this dictionary (or in another dictionary of the same type) for further use [15].

The advantages of this algorithm are that for each variant of the normal form it offers all the grammatical information (synthesized for non-existent words), this data can be used in the future to select one normal form from the set proposed by the program.

The disadvantages of this algorithm are that in the absence of the entered word, it cannot always cope with this task. He also does not cope with the diminutive form of the word.

Based on morphology, many scientists have developed a mathematical model of natural languages. For example, M.Kh. Khakimov in his works developed a mathematical model of a natural language [16-17].

Mathematical model - an approximate description of the object of modeling, expressed using mathematical symbols.

Mathematical models appeared along with mathematics many centuries ago. A huge impetus to the development of mathematical modeling was given by the emergence of computers. The use of computers made it possible to analyze and apply in practice many mathematical models that previously did not lend themselves to analytical research. A mathematical model implemented on a computer is called a computer mathematical model, and carrying out targeted calculations using a computer model is called a computational experiment.

The model can help explain the system and investigate the effects of various components, and make predictions about behavior.

The mathematical model of a natural language is a way to formally describe its syntactic and semantic constructions. The basis of syntactic constructions is the derivation of a word, and semantic constructions are the correct derivation of a phrase [16].

Morphological model of the Karakalpak language

The Karakalpak language belongs to agglutinative languages. “For agglutinative languages, one of the most productive ways of forming grammatical forms is affixation, that is, attaching grammatical particles-affixes to the root of the word, through which word formation or inflection is made” [18] For example: жала-клевета, жала+қор - slanderer, бала - child, бала+лық - childishness, пахта - cotton, пахта+шылық - cotton growing, бас - head, бас+қар - to manage, көп - a lot, көб+ей - to multiply.

There is a certain pattern in the joining of word formation affixes to the root or base of a word: the affixes of lexical and grammatical word formation, which form the main grammatical categories of parts of speech, are attached to the base first, then the affixes of functional grammatical word formation.

The affixes of inflection form various forms of the relationship of words to each other in a sentence. They make up four groups:

1) number affixes forming an intermediate form between word formation and inflection;
2) affixes of belonging, which serve to express the syntactic connection between the definition and the defined;
3) affixes of cases and postpositions that serve as formants to express the syntactic connection between an addition or circumstance, on the one hand, and the predicate, on the other;

face affixes expressing the subject-predicate connection of words in a sentence (балық+шы+лар+ы+мыз+да - in our fishermen) [19].

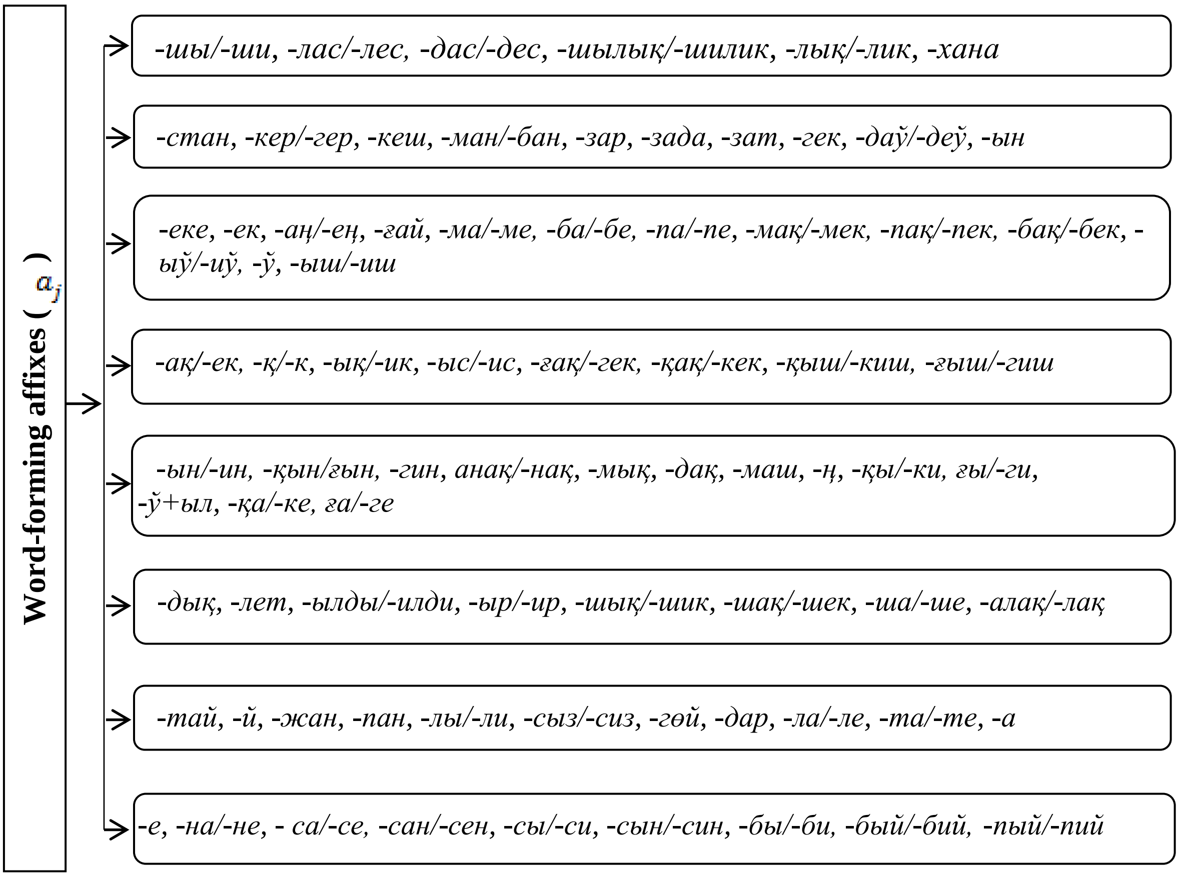
Word-forming affixes (a_j)
-шы/-ши, -лас/-лес, -дас/-дес, -шылық/-шилик, -лық/-лик, -хана
-стан, -кер/-гер, -кеш, -ман/-бан, -зар, -зада, -зат, -гек, -даў/-деў, -ын
-еке, -ек, -аң/-eң, -ғай, -ма/-ме, -ба/-бе, -па/-пе, -мақ/-мек, -пақ/-пек, -бақ/-бек, -ыў/-иў, -ў, -ыш/-иш
-ақ/-ек, -қ/-к, -ық/-ик, -ыс/-ис, -ғақ/-гек, -қақ/-кек, -қыш/-киш, -ғыш/-гиш
-ын/-ин, -қын/ғын, -гин, анақ/-нақ, -мық, -дақ, -маш, -ң, -қы/-ки, ғы/-ги, -ў+ыл, -қа/-ке, ға/-ге
-дық, -лет, -ылды/-илди, -ыр/-ир, -шық/-шик, -шақ/-шек, -ша/-ше, -алақ/-лақ
-тай, -й, -жан, -пан, -лы/-ли, -сыз/-сиз, -гөй, -дар, -ла/-ле, -та/-те, -а
-е, -на/-не, - са/-се, -сан/-сен, -сы/-си, -сын/-син, -бы/-би, -бый/-бий, -пый/-пий

Figure 1. Word-forming affixes that change the lexical meaning of a word in the Karakalpak language

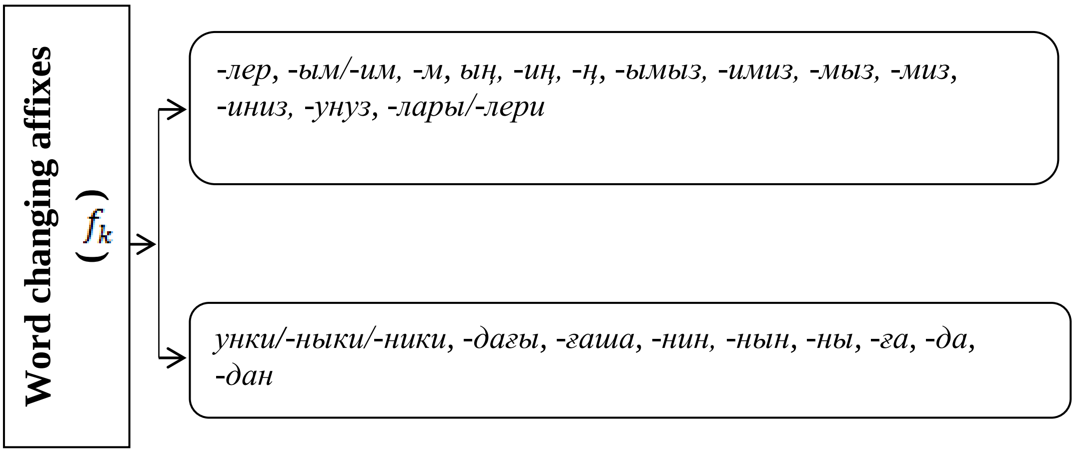
Word changing affixes
(f_k)
-лер, -ым/-им, -м, ың, -иң, -ң, -ымыз, -имиз, -мыз, -миз, -иниз, -унуз, -лары/-лери
унки/-ныки/-ники, -дағы, -ғаша, -нин, -нын, -ны, -ға, -да, -дан

Figure 2. Word changing affixes of the Karakalpak language

Figures 1, 2 show word-forming and inflectional affixes of the Karakalpak language.

Affixes are attached to the root of the word in a certain sequence (Fig. 3.):

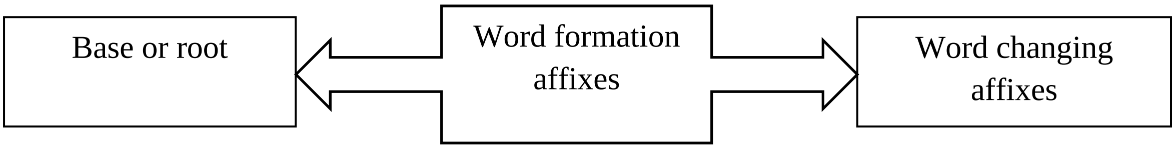

Figure 3. Scheme of creating complex words in the Karakalpak language

When one or another affix drops out, the order of agglutination is not violated, but before the addition of inflectional affixes, in the final result of word formation, the word must have a certain form: either a substantive or an attribute, since the inflectional affixes (plural, belonging, case and person) can be are attached only to certain functional-grammatical forms (acting in a sentence as a function of one or another member of the sentence).

Based on the circuit shown in Fig. 3 we have developed a mathematical model of complex word formation in the Karakalpak language.

$$C = \sum_{i=1}^{n} x_i + \sum_{j=1}^{m} a_j + \sum_{k=1}^{l} f_k$$

Here C – compound word, $\sum_{i=1}^{n} x_i$ – stem of words, where each word is a set *x*, $\sum_{j=1}^{m} a_j$ – word formation affixes, $\sum_{k=1}^{l} f_k$ – word changing affixes, where *i*, *j*, *k* – length of words and affixes.

Let's give an example based on the above scheme and a mathematical model for creating complex words (Fig. 4).

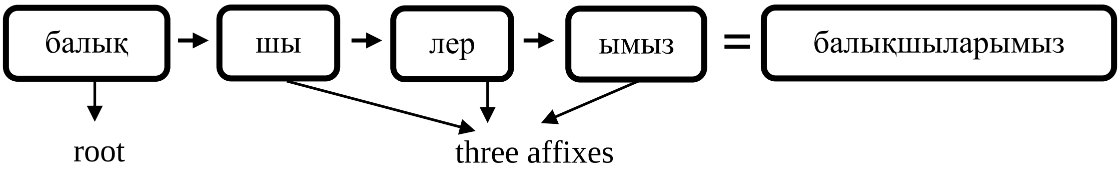

Figure 4. An example of using the word formation scheme in the Karakalpak language

1. «балық» - stem, root of a word.
2. «шы» - word formation affixes.
3. «лер» - word changing affixes.
4. «ымыз» - word changing affixes.

As you can see from the example, the word is formed by adding three words-affixes of the Karakalpak language to the root of the word.

Based on the above mathematical model, we will consider the sequence of composing complex words in the Karakalpak language.

$$C_1 = \begin{cases} \sum_{j=1}^{m} a_j(y_i) \sim I(z), & where\ z = y_0 \wedge y_1 \wedge y_2 \wedge \ldots \wedge y_{i-1} \\ \sum_{k=1}^{l} f_k(y_i) \sim I(z), & where\ z = y_0 \wedge y_1 \wedge y_2 \wedge \ldots \wedge y_{i-1} \end{cases}$$

$$C_2 = \begin{cases} \sum_{j=1}^{m} a_j(y_i) \sim I(z), & where\ z = y_0 \wedge y_1 \wedge y_2 \wedge \ldots \wedge y_{i-1} \\ \sum_{j=1}^{m} a_j(y_i) \sim I(z), & where\ z = y_0 \wedge y_1 \wedge y_2 \wedge \ldots \wedge y_{i-1} \end{cases}$$

$$C_3 = \begin{cases} \sum_{j=1}^{m} a_j(y_i) \sim I(z), & where\ z = y_0 \wedge y_1 \wedge y_2 \wedge \ldots \wedge y_{i-1} \\ \sum_{k=1}^{l} f_k(y_i) \sim I(z), & where\ z = y_0 \wedge y_1 \wedge y_2 \wedge \ldots \wedge y_{i-1} \\ \sum_{k=1}^{l} f_k(y_i) \sim I(z), & where\ z = y_0 \wedge y_1 \wedge y_2 \wedge \ldots \wedge y_{i-1} \end{cases}$$

$$C_4 = \begin{cases} \sum_{k=1}^{l} f_k(y_i) \sim I(z), & where\ z = y_0 \wedge y_1 \wedge y_2 \wedge \ldots \wedge y_{i-1} \\ \sum_{k=1}^{l} f_k(y_i) \sim I(z), & where\ z = y_0 \wedge y_1 \wedge y_2 \wedge \ldots \wedge y_{i-1} \end{cases}$$

Here C_1, C_2, C_3, C_4 – stem of words, I(z) – a single predicate, each word *z* is represented as a concatenation of two or more words $y_0 \wedge y_1 \wedge y_2 \wedge \ldots \wedge y_{i-1}$, where *i* – length of affixes.

In the Karakalpak language, we examined the cases of adding derivational and inflectional word affixes to the root word in 4 different forms.

1- step

C_1 = Word formation affixes + Word changing affixes.

2- step

C_2 = Word formation affixes + Word formation affixes.

3- step

C_3 = Word formation affixes + Word changing affixes + Word changing affixes.

4- step

C_4 = Word changing affixes + Word changing affixes.

Below is (Fig. 5) an example of the steps of the above diagram.

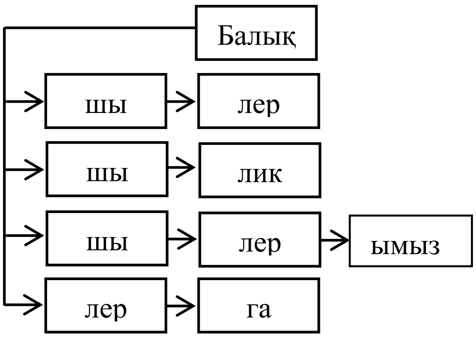

Figure 5. Scheme for creating compound words

Based on the above mathematical models, an algorithm for translating complex words has been developed as follows (Fig. 6).

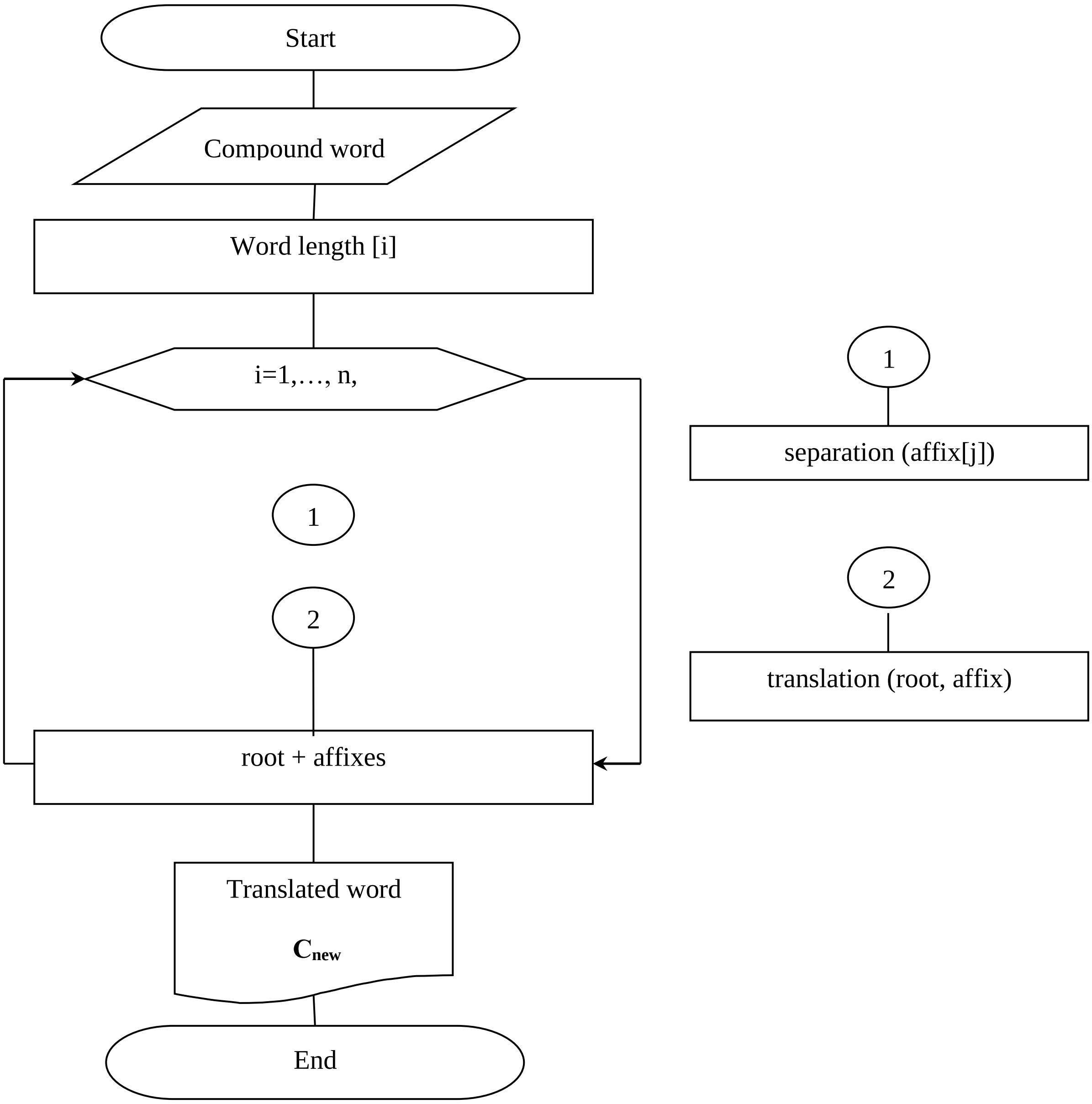

Figure 6. Algorithm that implements complex words

If we describe the flowchart step by step, the process of translating complex words will be done as follows:

Step 1. The compound word for translation is entered into the program.

Step 2. The program breaks words into parts.

Step 3. The stem and affixes are compared with the second language being translated.

Step 4. Once the comparison is done, translation is done by adding stems and affixes of the second language to be translated.

The above algorithm can be used to generate complex words for languages belonging to the Turkic language family and to complete the translation process.

Conclusion

The article deals with the morphology of the Karakalpak language, which belongs to the family of Turkic languages, and the order of word formation in the Karakalpak language. A mathematical model of complex word formation in the Karakalpak language has been developed by adding word-forming suffixes to the root word. On the basis of the developed mathematical model, an algorithm was developed for implementing the translation process in the Karakalpak language by dividing words into parts, that is, dividing them into stems and affixes. Since the Karakalpak language belongs to the family of Turkic languages, the developed mathematical model and algorithm can also be used in the process of translation between languages belonging to the family of other Turkic languages.

REFERENCES

[1]. Zubov A.V., Zubova I.I. Informatsionnye tekhnologii v lingvistike [Information Technologies in Linguistics]. Moscow, Akademiya Publ., 2004. 208 p.

[2]. 2. Marchuk Yu.N. Kompyuternaya lingvistika [Computer Linguistics]. Moscow, AST, Vostok-ZapadPubl., 2007. 320 p.

[3]. 3. Solovyeva A.V. Professionalnyy perevod s pomoshchyu kompyutera [Professional Translation by Means of Computer]. Saint Petersburg, Piter Publ., 2008. 160 p.

[4]. 4. Shevchuk V.N. Elektronnye resursy perevodchika [Electronic Translator's Resources]. Moscow, Librayt Publ., 2010. 136 p.

[5]. 5. Bowker L. Computer-Aided Translation Technology: A Practical Introduction. Ottawa, University of Ottawa Press, 2002. 185 p.

[6]. 6. Somers H., ed. Compute rs and Translation: A Translator's Guide. Amsterdam, Philadephia, John Benjamins Publishing Company, 2003. 349 p.

[7]. Kazennikov A.O. Postroyeniy morfologicheskogo analizatora neizvestnix slov na osnove slovarey sistemi ETAP-3 // Proceedings of the 34th Conference of Young Scientists and Specialists of IITP RAS "Information Technologies and Systems (ITiS'11)". Gelendzhik, 2011. p. 112-116.

[8]. Marchuk Yu. N. Problemi mashinnogo perevoda. Moscow: Nauka, 1983.201 p.

[9]. Bolshakov I.A. Uproshenniy morfologicheskiy analiz pri avtomaticheskoy proverke pravilnosti textov. // NTI, ser. 2, 1985, No. 6, p. 22-28.

[10]. Koval S.A. O sravnimosti i ekvivalentnosti kompyuternix predstavleniy morfologii // Computer linguistics and intellectual technologies. Tr. int. conference Dialogue'2003 (Protvino, June 11-16, 2003) / Ed. I. M. Kobozeva, N. I. Laufer, V. P. Selegeya. Moscow: Nauka, 2003. p. 305–311.

[11]. Creutz, Mathias and Lagus, Krista: Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. In: Publications in Computer and Information Science, Report A81, Helsinki University of Technology, (2005).

[12]. Zheltov P.V. Sravnitelno-sopostavitelniy analiz russkih, tatarskih i chuvashskih affiksov // Modern problems of science and education. - 2014. - No. 6.

[13]. Porter M. F. An algorithm for suffix stripping // Program. – 1980. – T. 14. – № 3. 130-137 p.

[14]. Sharipbaev A.A., Bekmanova G.T., Ergesh B.Zh., Buribaeva A.K., Karabalaeva M.Kh. Intellektualniy morfologicheskiy analizator, osnovanniy na semanticheskih setyah // Materials of the international scientific and technical conference "Open semantic technologies for the design of intelligent systems" (OSTIS-2012). Minsk, BSUIR, February 16-18, 2012 - 397-400 p.

[15]. Fedotov A.M., Sambetbaev M.A. Algoritm morfologicheskogo analizatora dlya kazahskogo yazika // Materials of the XVI All-Russian Conference. 2015 200-207 p.

[16]. Khakimov M.Kh. C modelyam yestestvennih yazikov dlya mnogoyazichnih situasiy kompyuternogo perevoda. Proceedings of the scientific conference "Problems of modern mathematics". April 22-23, 2011 Karshi. p. 531-538.

[17]. Khakimov M.Kh. Kop tilly computer tarjimasi uchun rus tiling semantics bazalari wa mathematician modellari. Problems of Informatics and Energy. 2011 no. 2.P.57-65.

[18]. Baskakov N. A. Karakalpakskiy yazik // II. Phonetics and morphology. Moscow: Publishing house of the Academy of Sciences of the USSR, 1952.544 p.

[19]. Berdimuratov Y., Allanazarov K., Patullaeva G. Karakalpak tili. Textbook for 5th grade. Nukus "Bilim" publishing house. 2015, 240 p.