The European Proceedings of
Social & Behavioural Sciences
EpSBS

Future Academy ISSN: 2357-1330

http://dx.doi.org/10.15405/epsbs.2017.12.3

ICPE 2017
International Conference on Psychology and Education

THE SOFTWARE SYSTEM FOR THE STUDY THE MORPHOLOGY OF THE KAZAKH LANGUAGE

Vladimir B. Barakhnin (a, b)*, Anatoliy M. Fedotov (a, b), Aigerim M. Bakiyeva (b), Murat N. Bakiyev (c), Saule Zh. Tazhibayeva (c), Tatiana V. Batura (b, d), Olga Yu. Kozhemyakina (a), Dzhamalbek A. Tussupov (c), Madina A. Sambetbaiyeva (b), Lyazzat Kh. Lukpanova (a, e)
*Corresponding author

(a) Institute of Computational Technologies SB RAS, 6 Lavrentiev Avenue, Novosibirsk, Russia, bar@ict.nsc.ru
(b) Novosibirsk State University, 2 Pirogova Street, Novosibirsk, Russia
(c) L.N.Gumilyov Eurasian National University, 2 Satpaev Street, Astana, Kazakhstan
(d) A.P.Ershov Institute of Informatics Systems SB RAS, 6 Lavrentiev Avenue, Novosibirsk, Russia
(e) The Kazakh National Research Technical University after K.I. Satpaev, 22a Satpaev Street, Almaty, Kazakhstan

Abstract

In the system of language teaching, one of the most important aspects is the study of the grammatical categories of lexical items. This process applies to the teaching of foreign languages, and to the learning the native language, both on basic and deeper level. The article describes the software system for studying the morphology of the Kazakh language. In the research process, 14 flexional classes were allocated for nouns and adjectives, 17 - for verbs. The dictionaries, including more than 5500 affixes and their combinations (taking into account the repetitions of combinations for different grammatical forms) were created. The quantitative volume of created dictionaries is sufficient to analyze the texts of any thematic affiliation. The system is supplemented by a dictionary of exceptions, including 18 nouns and 352 verbs, in which the basis is changing when the word is changing. During testing on words belonging to different parts of speech, there were no errors, that allows to judge about the correctness of the proposed algorithms. The system is equipped with a user-friendly web interface, allowing the learner to test his skills of word formation and of stemming of words of the Kazakh language.

Keywords: the Kazakh language, the study of morphology, word formation, stemming.

http://dx.doi.org/10.15405/epsbs.2017.12.3
Corresponding Author: Vladimir B. Barakhnin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

1. Introduction

In the system of language teaching, one of the most important aspects is the study of the grammatical categories of lexical items. This process applies to the teaching of foreign languages, and to the learning the native language, both on basic and deeper level. The understanding the language system is impossible without the understanding the variety of changes of word forms and their assignment to those or other categories of speech. In this regard, it is very important to create information-analytical systems with user-friendly interface that would allow, if necessary, to handle a substantial amount of the information and would let the student to visualize a system of the changes of word forms according to categories of speech that, in turn, presents opportunities for more effective development of the language system as a whole. This especially relates to the process of learning languages, which contain many of the branched categories of changes of word forms.

Of course, nothing can replace the intuitive perception of the subject of language as a system, his expert analysis and his understanding of the language in this conditional historical period of time. But it is impossible to deny the need to create the training systems which are capable, on the large-scale analysis of texts belonging to a particular language as the object of study, to introduce to the student the fullest possible picture of all grammatical categories and their implementations on the material of the study subject's language needed at this stage.

The modern theories of language learning suggest a wide use of the possibilities of the modern information technologies, including specially developed resources and programs that allow not only to accelerate the learning process, what is certainly one of the important arguments in their usage, but which help a student to form an idea of language as of complete system, the quantitative characteristics of which, in this way, the categories of language are serving. In turn, the scope of textual material, which can be processed by similar programs and resources, also contribute to the formation of student idea about the language in all its diversity that appears in the scale of test capabilities and permissible variations.

In addition, the process of language learning, whether learning the native for the subject or a foreign language, is inseparably linked with the development of literary space in all its diversity. To read and to understand a text in target language for the learner is a necessary component in the understanding of the practical use of the language. In this context the creation of systems that can process both the small volumes of text to identify, for example, phonetic features or characteristics of the rhymes (Hayward, 1996; Kaplan & Blei, 2007; Kao & Jurafsky, 2012; Kurt & Kara, 2012; Delmonte, 2013; Bobenhausen & Hammerich, 2015; Barakhnin et al., 2016), and the arrays of text to see the interdependence of such categories as genre, style, etc. (Barakhnin et al., 2017), are quite justified.

A practical point in learning the language, is undoubtedly one of the most important, and in this regard the undeniable importance of the application in learning process of the systems which allow the learner to see the implementation of categories of the target language in practice in real time, working with text and receiving the analysis of the proposed data. The interactivity of systems that process and analyze the proposed text, is an additional argument for the development of automated systems of analysis, of course, under condition of direction the obtained results to the highest possible accuracy, what, however, is a solvable problem.

http://dx.doi.org/10.15405/epsbs.2017.12.3
Corresponding Author: Vladimir B. Barakhnin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

2. Problem Statement

Let's consider the technology of creation of software systems to study the morphology of the language on the example of the Kazakh language. The Kazakh language is a Turkic language of Kipchak group, which refers to the type of synthetic, agglutinative languages and has a rich and complex morphology. As in other Turkic languages, the word consists of the basis to which the affixes are joined, expressing the different grammatical characteristics. To the basis of the word the several form-building affixes (sometimes called terminals) can join, and each such affixes executes the grammatical function which is inherent only for it, the order of the affixes is fixed.

3. Research Questions

The study of the morphology of the language assumes for both the perfect understanding of the advanced algorithms of word changes, and of the process of allocation of word basises called the stemming.

A distinctive feature of the proposed algorithms of the word-change is the usage of the principle of breaking words into morphological classes, each of them are characterized by a specific type of change of the alphabetic structure of word forms belonging to this class, in accordance with the principles outlined in the monograph (Belonogov & Novosyolov, 1979). The morphological classes of the Russian language are divided into two types: basis-changing and flexional, and for each variable part of speech there is, generally speaking, its division into morphological classes. The flexional classes are built on the base of analysis of syntactic function of words and the systems of their case, generic, personal, etc. affixes.

As for the task of the stemming, for its solution we use the well-known Porter's algorithm, which was published in 1980 for the English language (Porter, 1980). It described a sequence of steps, in each of which under certain rules, one of certain endings' transformations can happen. This rule has the following structure:

<condition> <ending > → <new end>.

The main idea of Porter's algorithm is that there are a limited number of form- and word-forming suffixes, and a word basis is converted without using any bases (the vocabularies) of basises: only an array of the existing suffixes (herewith the complex compound suffixes are broken down into simple) and the manually defined rules.

The fact that Porter's algorithm does not use any dictionaries and bases of basises, is a plus for performance and range of applications (it process rather good the non-existent words), and is a minus from the point of view of allocation of the basises. In addition, the disadvantage of the Porter’s algorithm is a frequently cited human factor: the rules for validation are manually set and are sometimes associated with grammatical features of the language, what increases the probability of error (Willett, 2006). However, this probability can be reduced by compiling a dictionary containing the words of exception.

4. Purpose of the Study

The purpose of this study is to create a software system for studying the morphology of the Kazakh language with a user-friendly web interface, allowing the learner to test their skills of word

http://dx.doi.org/10.15405/epsbs.2017.12.3
Corresponding Author: Vladimir B. Barakhnin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

formation and of the stemming of words of the Kazakh language. To achieve this goal, it's necessary to describe all flexional classes of nouns and verbs, to create the dictionaries of affixes and their combinations, and also the dictionary of exceptions, including words, in which a basis is changed when a word is changed.

5. Research Methods

5.1. The morphological model of the Kazakh language

In the Kazakh language, the word forms are formed by concatenation of the root and affixes (suffixes and endings). Each affix is associated with sets of morphological features, and the order of adding affixes is strictly defined. For example, for the nouns a suffix is added to the basis, an ending of the plural number follows, then a possessive ending, a case ending, and only then a personal ending (The grammar of Kazakh, 2002). The morphological features of adjectives used in the role of nouns (the adjectives are not changed in other cases) are analogous to the morphological features of nouns. Finally, for the verbs, first the negative ending is added to the basis, then a tense ending, then a personal ending (a negative ending and / or a personal ending may be absent).

New word forms are formed considering the morphological and semantic features of the initial forms by the following: first, the suffixes are added to the initial form of the word; then, moving from left to right, the category (voiceless, voiced, etc.) of the last letter (last sound) of the initial form of the word is determined for the addition of a specific ending (Bektayev, 1995).

The general formula for determining the composition of a word is the following:

түбір (root) +жұрнақ (suffix) + жалғау (ending)

The fundamental difference between the morphology of the Kazakh language and, for example, of the Russian morphology, is the presence in the Kazakh language (as in other Turkic languages) of the law of vowel harmony, according to which the affixes of a word are completely determined by the sound composition of its basis. Based on the analysis and grammar of the Kazakh language, the following basic rules of the Kazakh language can be determined (Valyaeva, 2017).

1. In the Kazakh language, the word cannot end on the voiced consonants*: б, в, г, ғ, д, ж*. In this case, there are some exceptions in which the suffix, which starts with the vowel, is removed, and the *б, г, ғ* at the end are transformed correspondingly into letters *п, к, қ*.

2. The softness and the hardness of words in the Kazakh language are determined by the presence of a certain vowel in the last syllable of the word. For example, a word is "hard" if there are vowels *а, о, ұ, ы, я* but it becomes "soft" if there are vowels *ә, ө, ү, і, е*. The hardness or the softness of words also correlates with the presence of some consonants: the word is hard if there are consonants *қ* and *ғ*, and soft if there are *к* and *г*.

3. After a hard syllable a hard ending follows, after a soft syllable a soft ending follows (for each morphological feature there are two forms of an ending, hard and soft).

4. Each next ending depends on the hardness of the previous one: if the last syllable of the word is hard, then each next ending will be hard, because the hardness of the next ending depends on the

http://dx.doi.org/10.15405/epsbs.2017.12.3
Corresponding Author: Vladimir B. Barakhnin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

previous one. Thus, if the word is hard, then all the endings are hard, and soft if the word is soft.

Formally for the nouns the following model of word formation is constructed. We denote by P_i the following types of endings (affixes) for $i = 1, 2, 3, 4$:

1) P_1 is a plural ending;
2) P_2 is a possessive ending;
3) P_3 is a case ending;
4) P_4 is a personal ending.

The following combinations of the noun endings are possible:

1) plural ending + possessive ending (P_1 P_2);
2) plural ending + case ending (P_1 P_3);
3) plural ending + personal ending (P_1 P_4);
4) plural ending + possessive ending + case ending (P_1 P_2 P_3);
5) plural ending + possessive ending + personal ending (P_1 P_2 P_4);
6) possessive ending + case ending (P_2 P_3);
7) possessive ending + personal ending (P_2 P_4);
8) case ending + personal ending (P_3 P_4).

For the verbs, there are the following types of endings:

1) P_1 is a negative ending;
2) P_2 is a tense ending;
3) P_3 is a personal ending.

The following combinations of verb endings are possible:

1) tense ending (P_2);
2) tense ending + personal ending (P_1 P_3);
3) negative ending + tense ending (P_1 P_2);
4) negative ending + tense ending + personal ending (P_1 P_2 P_3).

5.2. The flexional classes of nouns, adjectives and verbs of the Kazakh language

The base for the constructions of the algorithms for morphological analysis and synthesis is the division of all words into classes that determine the character of the change in the literal composition of word forms. These classes are conditionally called morphological. The changes in the forms of the words can have a different character. They can be related both to the change of the forming affixes of the word and to its basis (what is extremely rare in the Kazakh language: for example, there are 18 exceptions for

http://dx.doi.org/10.15405/epsbs.2017.12.3
Corresponding Author: Vladimir B. Barakhnin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

nouns, and 352 for verbs).

In the process of investigation of the structured rules for attachment of endings given in (The grammar of Kazakh, 2002; Bektayev, 1995; Valyaeva, 2017), we established 14 flexional classes for the nouns of the Kazakh language:

1) hard word, the basis ends with a vowel (except *у*);
2) soft word, the basis ends with a vowel (except *у*);
3) hard word, the basis ends with *б, в, г, д*;
4) soft word, the basis ends with *б, в, г, д*;
5) hard word, the basis ends with *ж, з*;
6) soft word, the basis ends with *ж, з*;
7) hard word, the basis ends with *л*;
8) soft word, the basis ends with *л*;
9) hard word, the basis ends with *м, н, ң*;
10) soft word, the basis ends with *м, н, ң*;
11) hard word, the basis ends with *р, у, й*;
12) soft word, the basis ends with *р, у, й*;
13) hard word, the basis ends with a dull consonant;
14) soft word, the basis ends with a dull consonant;

For verbs we established 17 inflectional classes:

1) hard word, the basis ends with a vowel (except *ю*);
2) soft word, the basis ends with a vowel (except *ю)*;
3) hard word, the basis ends with *б, г, ғ*;
4) soft word, the basis ends with *б, г, ғ*;
5) hard word, the basis ends with *з*;
6) soft word, the basis ends with *з;*
7) hard word, the basis ends with *р, л*;
8) soft word, the basis ends with *р, л*;
9) hard word, the basis ends with *м, н, ң*;
10) soft word, the basis ends with *м, н, ң*;
11) hard word, the basis ends with *ж, д*;
12) soft word, the basis ends with *ж, д*;
13) hard word, the basis ends with a dull consonant;
14) soft word, the basis ends with a dull consonant;
15) hard word, the basis ends with *ю*;
16) soft word, the basis ends with *ю*;
17) hard word, the basis ends with *у*.

http://dx.doi.org/10.15405/epsbs.2017.12.3
Corresponding Author: Vladimir B. Barakhnin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

The listed partition of the words into flexional classes completely and without any crossing covers all possible variants of the words of the Kazakh language, what means a complete correct solution of the task (it can be noted that some sub-variants are not realized: for example, hard words can not end with the letter г, but we have indicated also this combination, so we had the formal completeness of the coverage). The dictionaries, including more than 5500 affixes and their combinations (the variants of endings) for 14 flexional classes of nouns and adjectives, and also about 2000 verb affixes and their combinations for 17 flexional classes (some combinations of the affixes are repeated), were created. These dictionaries are used in the software application for the generation of the word forms, which will be described below.

5.3. The algorithm of the stemming of the word forms

As already noted, in the basis of our implemented algorithm of the stemming of words of the Kazakh language is the Porter's algorithm. Depending on the execution of the criterions the decision is taken, whether the basis of word was got or it is required the truncation of an affix. The algorithm for obtaining the basises consists of the following steps.

1. The input is data is any word form (verb, noun, adjective).
2. Starting with the last letter of the word the search is occurred in the list of affixes.
3. If this affix is found, it is separated. The remaining part of a word after the truncation of all affixes is the basis.

The main problem of the described algorithm is the presence in the Kazakh language of the words where the last letters of the basis are matched with one of the affixes. In this case, the algorithm can cut off more than necessary. The only possible mechanism to prevent such errors, is the compilation of a dictionary of basises, the last letters of which match the one of the affixes.

It should be noted that the proposed algorithm of the stemming (as, indeed, of the generation) is applicable only to simple forms of the verbs. The more complex verb forms consisting of 2-4 words, are planned to consider in the future. However, in scientific and technical texts the complex verbs are not used.

6. Findings

Based on the described algorithms we developed a web application generation and stemming of the word forms of nouns, adjectives and verbs in the Kazakh language, which is in the public domain in the Internet (Bakieva, 2017). The recommended browsers are Google Chrome, Mozilla FireFox. The web application demonstrates the principal features of the system. Although the web interface slows down the work (the word forms of a given word are generated in about 25 seconds), however, the actual implementation of our software in its integration with linguistic software systems allows to get the results faster: the time to generate all word forms of particular words is 1 second. It is assumed to refer to the current version and not to the web application. The generation module and the module of stemming are

http://dx.doi.org/10.15405/epsbs.2017.12.3
Corresponding Author: Vladimir B. Barakhnin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

implemented in Python using the libraries: psycopg2, collections. The dictionaries are stored in a PostgreSQL database.

For the practical verification of the skills of word formation upon request of the student:

a) a basis is randomly selected from the database of the basises,

b) a type is randomly allocated from the table of types of word changes (number, case, etc.)

c) the system offers an interface that allows to enter, for example, all cases.

The learner generates all of the required forms, and enters them into the system, the system determines whether each word form is identified correctly or incorrectly, in case of an error the system gives the correct answer (fig. 01).

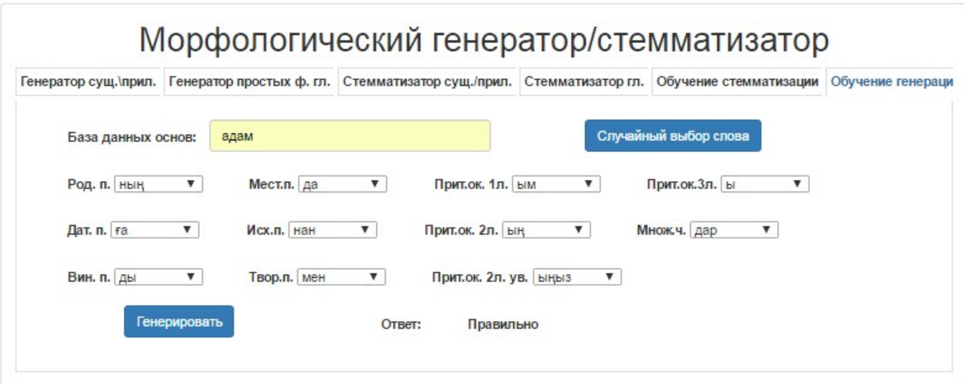

Figure 01. The interface of the test the skills of word formation

Similarly, for the practical testing of skills of the stemming at the request of the student:

a) a basis is randomly selected from the database of the basises,

b) an ending is randomly selected from a database of combinations of affixes for a given flexional class,

c) finally the word form of the given word is generated and offered for the student for the stemming.

The student makes the stemming of a word, enters the extracted basis in the system, the system determines whether the basis is determined right or wrong, in the case of an error the system gives the correct answer (fig. 02).

http://dx.doi.org/10.15405/epsbs.2017.12.3
Corresponding Author: Vladimir B. Barakhnin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

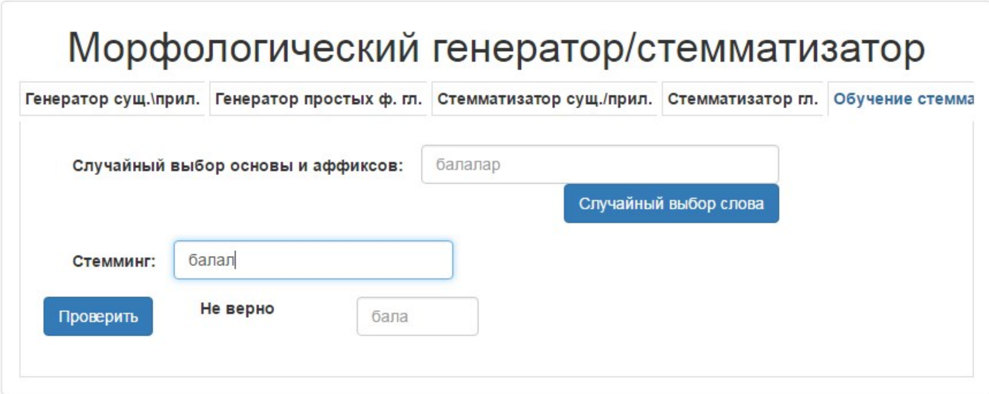

Figure 02. The interface of the test the skills of the stemming

7. Conclusion

The software system for studying the morphology of the Kazakh language is described. In the research process, 14 flexional classes were allocated for nouns and adjectives, 17 – for verbs. The dictionaries, including more than 5500 affixes and their combinations (taking into account the repetitions of combinations for different grammatical forms) were created. The quantitative volume of created dictionaries is sufficient to analyze the texts of any thematic affiliation. The system is supplemented by a dictionary of exceptions, including 18 nouns and 352 verbs, in which the basis is changing when the word is changing. During testing on words belonging to different parts of speech, there were no errors, that allows to judge about the correctness of the proposed algorithms.

The system is equipped with a user-friendly web interface, allowing the learner to test their skills of word formation and lemmatization words of the Kazakh language.

Acknowledgments

Work is executed with partial support of the Presidential programme «Leading scientific schools of RF» (grant 7214.2016.9).

References

Bakieva, A. M. (2017). The program of the generation and of the stemming of the word forms of the Kazakh language. Retrieved from http://db4.sbras.ru/morpher [In Russian].

Barakhnin, V B., Kozhemyakina, O. Yu., Zabaykin, A. V. (2016). Usage of modern computer technologies in the learning process of the philologists of complex analysis of Russian poetic texts. *SHS Web of Conferences, 29, 02002* http://dx.doi.org/10.1051/shsconf/20162902002

Barakhnin, V., Kozhemyakina, O., Pastushkov, I. (2017). Automated determination of the type of genre and stylistic coloring of Russian texts. *ITM Web of Conferences 10(02001)* https://doi.org/10.1051/itmconf/20171002001

Bektayev, K. (1995). The Large Kazakh-Russian, Russian-Kazakh dictionary. Almaty. [In Russian].

Belonogov, G. G., Novosyolov, A. P. (1979). The automation of processes of storage, search and generalization of the information. Moscow: Nauka. [In Russian].

http://dx.doi.org/10.15405/epsbs.2017.12.3
Corresponding Author: Vladimir B. Barakhnin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

Bobenhausen, K., Hammerich, B. (2015). Métrique littéraire, métrique linguistique et métrique algorithmique de l'allemand mises en jeu dans le programme Metricalizer². *In: Langages, 199, Traitement automatique des textes versifiés: problématiques et pratiques,* 67-87.

Delmonte, R. (2013). Computing poetry style. *In: C.Battaglino, C.Bosco, E.Cambria, R.Damiano, V.Patti, P.Rosso (eds.). Proceedings of 1st International Workshop ESSEM 2013, CEUR Workshop Proc., 1096,* 148-155.

Hayward M. (1996). Analysis of a corpus of poetry by a connectionist model of poetic meter. *Poetics, 24(1),* 1-11.

Kao, J., Jurafsky, D. (2012). A computational analysis of style, affect, and imagery in contemporary poetry. *NAACL Workshop on Computational Linguistics for Literature.* http://web.stanford.edu/~jurafsky/kaojurafsky12.pdf

Kaplan, D. M., Blei, D. M. (2007). A computational approach to style in american poetry. *7th IEEE International Conference on Data Mining (ICDM 2007).* 553-558.

Kurt, A., Kara, M. (2012). An algorithm for the detection and analysis of arud meter in Diwan poetry. *Turkish journal of electrical engineering & computer sciences, 20(6),* 948-963.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: Electronic Library and Information Systems, 14(3),* 130–137.

The grammar of Kazakh. Phonetics, word formation, morphology, syntax. (2002) Astana: Astana-poligraphy. [In Kazakh].

Valyaeva, T. (2017). Grammar of the Kazakh language. Retrieved from http://kaz-tili.kz/ [In Russian].

Willett, P. (2006). The Porter stemming algorithm: then and now. *Program: Electronic Library and Information Systems, 40(3),* 219–223.