Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

THE APPLICATION OF STATISTICAL METHODS IN THE DEVELOPMENT OF CYRILLIC-LATIN CONVERTER FOR TATAR LANGUAGE

A. V. Danilov[1,*], L. L. Salekhova[1], N. Anyameluhor[2]

[1]Kazan Federal University, Institute of Philology and Intercultural Communications

[2]Nottingham Trent University (Great Britain), Department of Computing and Technology

Published online: 24 November 2017

ABSTRACT

The article describes the process of a software product development that allows you to convert a text written in Tatar to Latin using Cyrillic graphics. The aspects of Cyrillic graphics to Latin graphics conversion are considered for Tatar language. The authors study the application of various statistical methods necessary for converter operation and analyze the speed and the accuracy of the conversion algorithms.

An algorithm was created and software modules were developed that made it possible to convert messages written in Tatar Cyrillic alphabet to Tatar Latin alphabet. Based on normative documents and scientific works on the use of Latin graphics in Tatar language, a verbal and an algorithmic model of conversion was constructed. In the process of development, it turned out that the process of a Tatar word conversion depends on its origin. If native Tatar words are converted according to the phonetic principle (кәләм - qäläm), the borrowed words are converted according to the rules of transliteration. The main problem of the study is the problem of a word origin determination. In order to solve this problem, the authors propose various algorithms. Software tools based on the statistical processing of linguistic data are considered and developed in the work: combined bigram analysis, naive Bayesian classification and a direct search. Each of these algorithms is used to determine the etymology of a word, on which depends the application of certain rules of conversion from Cyrillic to Latin.

Author Correspondence, e-mail: tukai@yandex.ru

doi: http://dx.doi.org/10.4314/jfas.v9i7s.106

The result of the research is a developed software product that is capable to carry out the process of Cyrillic graphics conversion to Latin for Tatar. In the future, the authors plan to improve the software product and use it in educational activities.

INTRODUCTION

Various projects are being developed and implemented in the Republic of Tatarstan (RT), the purpose of which is the use of information and communication technologies for the development of Tatar language and culture. The search engine Google began to support the search for information in Tatar language. The Tatar version of the operating system Windows XP, Vista, 7 has been developed. The portal of the Republic of Tatarstan Government presents information in two official languages of the republic - Tatar and Russian [1,2,3]. An English-Russian-Tatar dictionary of computer terms has been published, in which more than 7,000 computer terms have been translated into Tatar language. A universal encoding Unicode - the most common encoding - has a set of Tatar symbols by default.

Also the works to support the use of Tatar language for different alphabets are performed.

During the history of its development Tatar language changed its writing several times [4].

• Arabic graphics was used before 1927. It is worth noting that the Tatars living in China still use arabica.

• In 1927-1939, the Latin alphabet was used. Currently, the Latin alphabet is used by the Tatars living in Turkey, Finland, the Czech Republic, Poland, the United States and Australia.

• From 1939 to the present time Cyrillic is used, adapted for Tatar language by the addition of six letters. Baptized Tatars has been used Cyrillic since the XIXth century.

In 2012, the government of the Republic of Tatarstan adopted the law "On the use of Tatar language as the state language of the Republic of Tatarstan" [5]. According to this document, it is possible to use three alphabets for the writing in Tatar language - Cyrillic, Latin and Arab one.

The use of alternative sign systems has its advantages. The alphabet based on the Latin alphabet Yanalif, which was adopted in the thirties of the 20th century, allowed Tatar-speaking pupils to learn European languages more easily, expanding the possibilities of communication with other Turkic-speaking ethnic groups who also began to use the Latin alphabet in their writing and continue to do this even now.

In conditions with a parallel use of two alphabets - Latin alphabet and Cyrillic alphabet - there is a need to convert messages from one graphic to another. The relevance of the research

aimed at the conversion automation of a message written in Tatar Cyrillic alphabet to the Tatar alphabet is not doubted.

METHODS

The main objective of the study is to design an algorithmic model for conversion and the development of a software product on its basis that allows you to convert the messages written in Tatar language into the Latin script using Cyrillic alphabet.

The development of the converter model was based on the rules of Latin graphics use in Tatar language and the transition from Cyrillic to Latin, set out in the Law of RT "On the use of Tatar language as the state language of the Republic of Tatarstan" dated on 24.12.2012. It regulates the use of Tatar language in three versions - in Cyrillic, Latin and Arabic. We also used the rules of translation and the use of Latin graphics proposed by V. Khakov in the work «Телен белгән ил ачар: Латин графикасында уку һәм язу кунекмәләре» [6]. According to V.Khakov, most of the letters of the Tatar Cyrillic alphabet are converted unequivocally. However, there is a number of letters, the conversion of which depends on the specific use of a considered letter in words. The presence of these exceptions led us to the need of developing the principles for conversion and taking into account the context of a letter use in different kinds of words.

The principle is based on a word etymology, that is various rules of transliteration are used depending on the origin of a word.

A set of rules is applied to the words of Tatar, Arab and Persian origin, based mainly on the phonetic principle that is, we write as we hear - тавык - tawıq.

A set of rules is applied to borrowed words, mainly of Russian and English origin, which is close to mechanical transliteration, that is, a mutually ambiguous correspondence between a symbol on the Cyrillic alphabet and a symbol in Latin - кавалерия - kavaleriya.

Based on this principle, a simplified conversion algorithm is compiled, which is presented in the flowchart (Fig. 1):

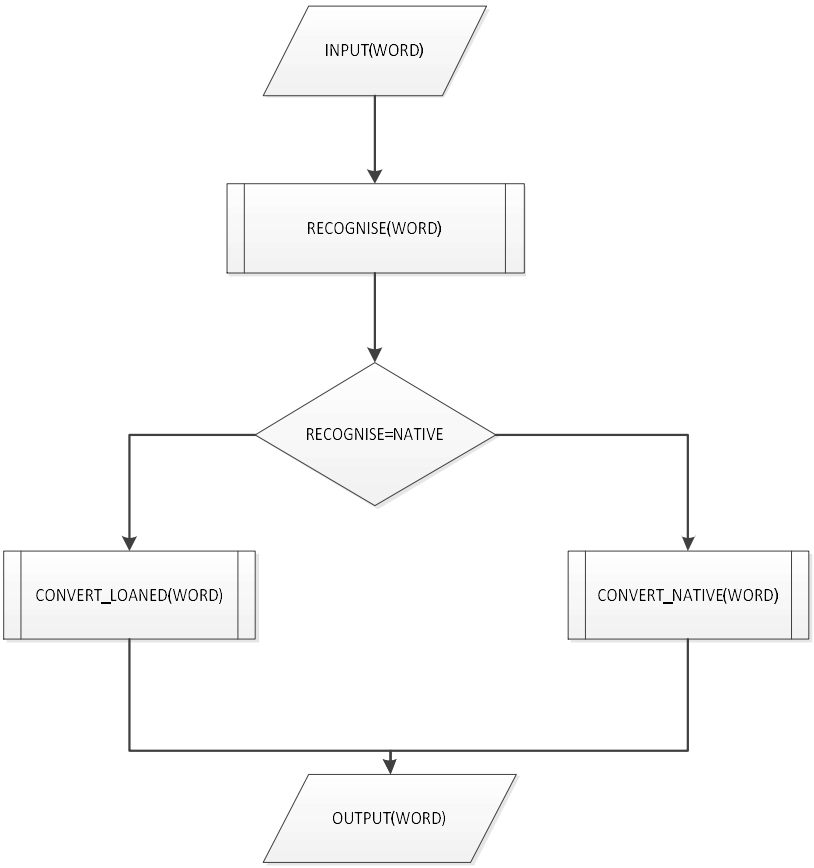

Fig.1. Algorithm of a message translation written in Tatar language to Latin script using Cyrillic graphics

The input of the program is provided with a text message for conversion, then the processing takes place according to the units developed in the algorithm:

- **INPUT(WORD)** – a Tatar word, written in Cyrillic is supplied to the program input (WORD);
- **RECOGNISE(WORD)** – the detection of a word etymology (a native or a borrowed one);
- **RECOGNISE=NATIVE** – the conditional selection unit, depending on a test result, the word is converted according to a certain set of rules:

1st case **(CONVERT_NATIVE)** - if a word is of Tatar, Arabic and Persian origin, then it is converted by character. This unit contains a set of procedures and functions intended for conversion. All procedures are built taking into account the rules of translation from Cyrillic to Latin.

2nd case **(CONVERT_LOANED)** - if a word is a borrowed one, it is converted according to the conversion rules for borrowed words symbol by symbol. This unit contains a set of procedures and functions that implements the process of mechanical transliteration.

• **OUTPUT (WORD)** - a word is displayed recorded in the Latin alphabet.

The conversion process is not time consuming and it is programmed easily. The most laborious task is to develop the methods for word etymology detection necessary for further conversion.

After the analysis of possible options, the following methods of a word etymology detection were chosen: combined bigram analysis, direct search method and naive Bayes classifier.

The method of direct search (brute force method)

When this method of etymology detection is applied, an original word is checked for the consistency from a preliminary prepared body of borrowed words.

The disadvantage of the algorithm is that it will work only if a word enters a body. If the a is absent in a body, then the results of the analysis may be contradictory ones.

Combined bigram analysis

The name of this method was obtained due to the combination of two technologies for a word etymology detection - bigram analysis and morphological analysis.

The principle of the bigram analysis operation is the statistical analysis of bigrams (pairs of letters that are near), forming a word. Some bigrams are more common in the language than others, therefore, it is possible to determine the origin of a word using statistical methods, having analyzed the bigrams included in it. Similar algorithms are widely used in the web industry [7,8]. In particular, in Internet browsers to determine the encoding of a web page. Internet company Yandex uses similar methods to determine an automatically generated text on web pages. The authors developed a special program to identify bigrams, which analyzes texts that consist exclusively of native Tatar words.

The bigram analysis without modifications has a significant drawback. Tatar language is agglutinative, i.e. the dominant principle of word formation is agglutination - the "gluing" of new morphemes to the end of a word. Often there is a situation when a borrowed word with a short root morpheme has a long suffix part. The bigrams, presented in suffix morphemes, are used in Tatar language very often. When an abovementioned algorithm of bigram analysis is used, the program will define this word as a native one. Therefore, it is necessary to select the root part of a word during an analysis, after which only the root can be analyzed.

In order to solve this problem, it was decided to use a mechanism that makes it possible to "cut off" a root morpheme from a suffix one, and analyze only the bigrams entering the root.

For these purposes researchers develop special programs-stemmers, but such a program is not developed for the Tatar language. Thus it was decided to use a morphological analyzer for Tatar language, developed by the Institute of Applied Semiotics at the Academy of Sciences of the Republic of Tatarstan.

Naive Bayes classifier (NBC)

This method is similar to CBA, however in this case the detection of a word etymology is reduced to the solution of classification and the application of the naive Bayesian classifier issues. The method is very powerful and at the same time universal one, it has found application in many areas of IT industry. For example, it is used to protect against spam - a classifier based on a dictionary loaded in it and frequency characteristics helps to catch letters containing advertisements and prevents their appearance in a user's mailbox [9,10].

All three methods of etymology detection were implemented in the software product.

In order to implement the combined bigram analyzer and the naive Bayes classifier, a number of auxiliary components and resources was developed. First, the training corps of native and borrowed words was collected, on the basis of which analysis and classification were made. Secondly, since both algorithms work with statistical data about bigrams, an algorithm has been created to split a text into bigrams. The algorithm allows you to represent all text in the form of a set of bigrams and automatically calculates their frequency characteristics. Thus, using the data on the frequency characteristics of bigrams presented in the cases, the program builds a probabilistic model for a combined bigram analyzer and a naive Bayesian classifier. Thirdly, the module was developed to exchange data with a morphological analyzer located on AS RT server.

Let's consider a simplified procedure for both detection methods (RECOGNIZE unit in the flowchart (Figure 1)).

1. The program refers to the morpheme analyzer for the analysis of a Latinized word;
2. The morpheme-analyzer marks a root morpheme and an affix;
3. A root morpheme is divided into bigrams;
4. The resulting set of bigrams is analyzed using CBA/NBC;
5. The word is classified as a native/a borrowed one.

The program implemented both detection methods. It is possible to choose CBA and NBC for latinisation.

In order to implement the definition of word etymology, a body of borrowed words was compiled by direct enumeration method. The volume of the corpus makes about 300,000

word forms, mainly consisting of Russian-language borrowings. The morphological analyzer is used to increase the accuracy of a work.

RESULTS

After the analysis of the diagram unit components, software development began. A team of developers and testers was created, consisting of staff and students from the Institute of Philology and Intercultural Communication of KFU. The software was developed using the integrated development environment Microsoft Visual Studio 2015.

The appearance and the interface of the preliminary version of the program is shown on Fig. 2.

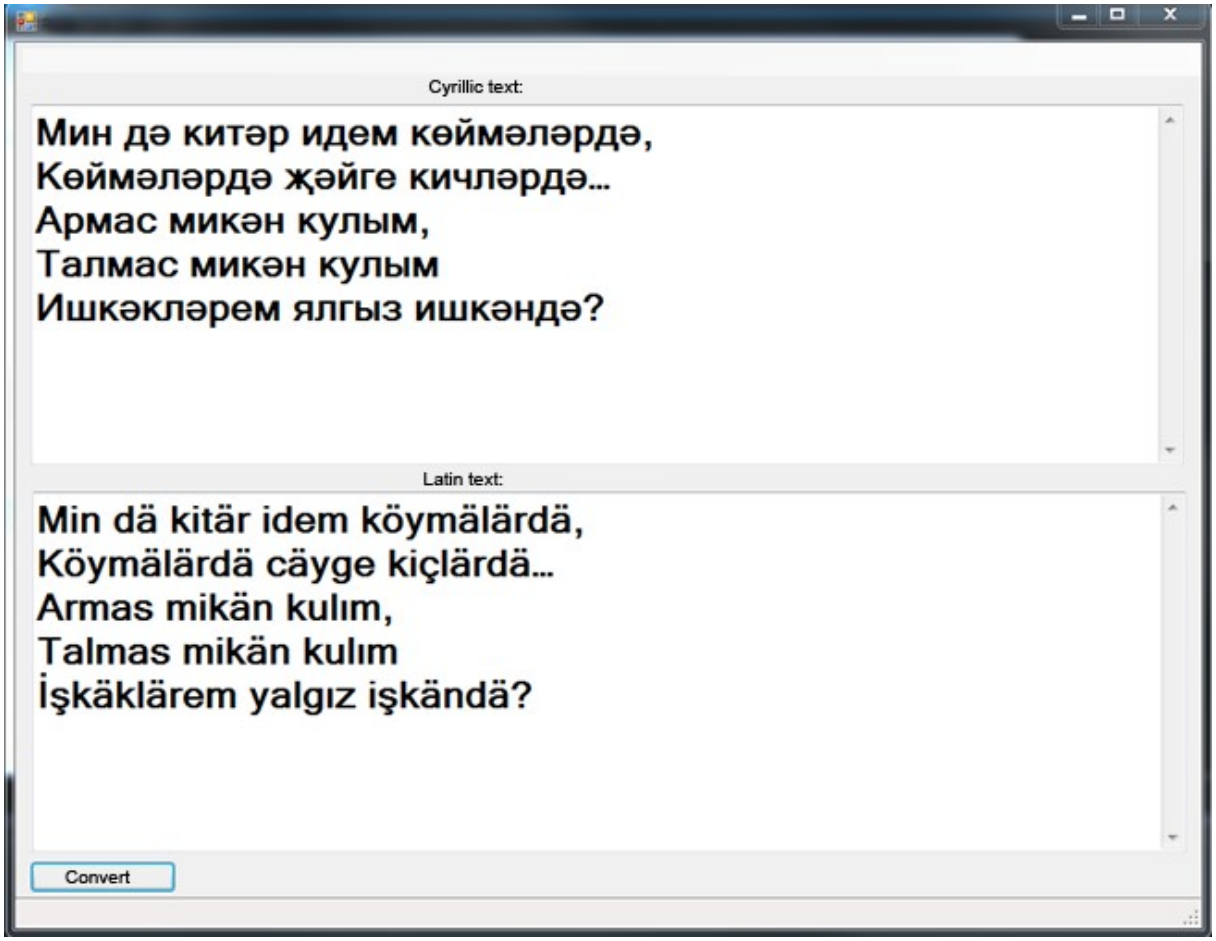

Fig 2. Program interface

Then they implemented the procedure to test and compare the speed and the accuracy of the detection methods discussed in this article. It was necessary to determine the parameters of the program to obtain an optimal result. To do this, we took a test sample consisting of 1400 words, which were processed by the program. A test sample included both native and borrowed words.

DISCUSSION

The analysis of the obtained data is presented in Table 1.

Table 1. Program operation data for a test sample of words processing.

Etymology detection method	CBA	NBC	Direct search
Amount of sample and words	1396		
Number of words with a correctly defined etymology	791	1083	1041
Percentage of total sample size	56,7 %	77,6 %	74,6 %
Work speed, sec	834	830	849
Number of convertible words per second	1,67	1,68	1,64

The obtained data showed that the most accurate detection method is the naive Bayes classifier, which allows to obtain fairly accurate results. Nevertheless, the speed of each of the methods leaves much to be desired. A slow conversion is associated, first of all, with the exchange of data between a program and a morpheme analyzer. It was decided to check the speed and the accuracy of the program without an analyzer evaluation (see Table 2).

Table 2. Program operation data for a test sample of words processing without a morphological analyzer use

Etymology detection method	CBA	NBC	Direct search
Amount of sample and words	1396		
Number of words with a correctly defined etymology	843	1058	1092
Percentage of total sample size	60,4 %	75,8 %	78,2 %
Work speed, sec	160	168	175
Number of convertible words per second	8,7	8,3	7,9

SUMMARY

The obtained results indicate that if an analyzer refuses to use a morpheme analyzer, the speed of the converter operation increases 4 times without a significant damage in conversion accuracy. Using the direct search method, the highest accuracy is achieved - 78% (in this case, the direct search method was used).

Thus, it can be concluded that the use of the morpheme analyzer slows down the program, and the rejection of it does not adversely affect the accuracy of data processing.

ACKNOWLEDGEMENTS

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

REFERENCES

1. Internet resource: Information materials on the final results of All-Russia Population Census of 2010 / Federal Service of State Statistics website, 2010, URL: http://www.gks.ru/free_doc/new_site/perepis2010/perepis_itogi1612.htm (reference date: 01/22/2017)

2. A.Danilov and L.Salekhova. Design of Virtual Keyboard for Tatar-Speaking Users on the Basis of the Mobile Operating System Android/ International Journal of Soft Computing, №10 (5): pp. 348-352, 2015.

3. Zaripova R.R. L.L. Salekhova, N.K. Tuktamyshov, R.F. Salakhov. Definition of development level of communicative features of mathematical speech of bilingual students / //Life Science Journal. – 2014. – . №11(8). – URL: http://www.lifesciencesite.com/lsj/life1108/ (reference date: 23.03.2017)

4. Tatar writing. – URL: https://ru.wikipedia.org/wiki/Татарская_письменность (reference date: 10.03.2016)

5. On the use of Tatar language as the state language of the Republic of Tatarstan [Electronic resource]: the law of the Republic of Tatarstan No. 1-ZRT issued on January 12, 2013 - Access mode: http://mon.tatarstan.ru/rus/file/pub/pub_227812.pdf

6. Khakov V.Kh. Теленбелгэн ил ачар: Латинграфикасындауку һәм язу күнекмәләре[Text]/ Khakov V.I. - Kazan: Mgarif Publishing House, 1993 - 140 p.

7. E.A. Grechnikov, G.G. Gusev, A.A. Kustarev, A.M. Raygorodsky. Search for unnatural texts // Proceedings of VLDB-2001, 2001, 306-308

8. J. Attenberg, T. Suel. Cleaning search results using term distance features // Proceedings of AIRWeb-2008, pp. 21-24

9. A. Benczur, I. Biro, K. Csalogany, and T. Sarlos. Web spam detection via commercial intent analysis. //Proceedings of AIRWeb-2007, pp. 89-92, New York, NY, USA, 2007

10. C. Manning. Foundations of Statistical Natural Language Processing // The MIT Press. – 1999. – 364 P.

How to cite this article:
Danilov A V, Salekhova L L, Anyameluhor N, . The application of statistical methods in the development of cyrillic-latin converter for tatar language. J. Fundam. Appl. Sci., 2017, *9(7S), 1174-1183.*