THE DESIGN OF CYRILLIC-LATIN CONVERTER FOR TATAR LANGUAGE

Andrey V. Danilov*, Leila L. Salekhova, Bulat E. Khakimov

Kazan Federal University, 420008, Kazan, Tatarstan Str., 2 **(RUSSIAN FEDERATION)**
***Corresponding author:** tukai@yandex.ru

DOI: 10.7813/jll.2016/7-2/52

Received: 05 Apr, 2016
Accepted: 15 May, 2016

ABSTRACT

The relevance of the study is conditioned by the need of computer program development which allow to transliterate texts from Cyrillic graphics during the operation in non-Russian systems. During transliteration a text written with the use of a particular alphabet, is presented by the alphabet of another system. The correspondence of only two alphabet letters is usually taken into account. However, during the transliteration of living languages they try to take into account the sound aspect in order not to detach a word from its live-sounding form. Thus, they transliterate not an alphabet, but the graphics adopted in this language system. Modern Tatar language has Cyrillic graphics. In this regard, the article shows the process of software development, which allows you to convert a text written in Tatar language using Cyrillic symbols into the Latin symbols. The principle of conversion is proposed, based on etymology. Original Tatar words are proposed to convert according to phonetic principle, and the borrowed words are proposed to convert according to transliteration rules. In order to determine the origin of a Tatar word it is proposed to use the following algorithms: digram analysis, combined analysis and search. An algorithmic model of conversion was developed. The software designed by the authors allows to transliterate native Tatar words nowadays.

Key words: transliteration, Cyrillic, Latin, converter, Tatar, Russian

1. INTRODUCTION

Writing is one of communication means for a man. The need for communication at a distance led to the emergence of writing, it expanded the communication circle and united people not only in space but also in time. The invention of writing led to the information revolution, which provided new opportunities for the exchange and the transfer of information. The writing and reading skills are the necessary conditions for learning. Despite the emergence of modern data transfer technologies, written communication has not lost its significance today.

Throughout its history, Tatar language changed its script several times. Until 1927 Arabic script was used, from 1927 to 1939 the Latin script was used, since 1939 to the present time the Cyrillic script is used. During the thirties of the 20-th century after the adoption of the alphabet "Yanalif" many books were reprinted using the Latin alphabet. According to V.G. Hakov (Hakov V.G., 1993), the practice of those years showed that the use of the Latin alphabet for the Tatar script enabled to facilitate the assimilation of European languages, the ability to read books written in Turkic languages.

Modern Tatarstan is a multinational region of Russian Federation, where the representatives of various nationalities and cultures live. According to the National Population Census of 2010, the majority of the population of the Republic of Tatarstan is presented by Tatars (53.15%) and Russians (39.1%) [2]. In accordance with the law "On the use of Tatar language as a state language of the Republic of Tatarstan" the official languages of the Republic of Tatarstan are Tatar and Russian [3]. Thus, the number of texts of different styles (formal, business, scientific, artistic, journalistic, colloquial ones) written in Tatar language is being increased. The functionality of Tatar language, as the language of storage, processing and transmission of information in the field of computer technology is being increased [Nevzorova O., Suleymanov D., Khakimov B, 2013; Danilov A, Salekhova L., 2015]. There are many transliterated words from Russian language, but the transliterators for Tatar language are not developed. This fact explains the relevance of our development.

2. METHODS

The purpose of this study was to design an algorithmic conversion model and develop a software on its basis, which allows to "translate" the messages recorded in Tatar into the Latin alphabet using Cyrillic one.

The basis of transliterator model development consisted of Latin script use rules in Tatar language and the transition from the Cyrillic to the Latin alphabet, set out in the Law "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 the Tatar language in three versions: Cyrillic, Latin and Arabic [3]. Also, the translation rules and the application of Latin alphabet

were used, proposed by V. Hakov in the work "Теленбелгән ил ачар: Латин графикасында уку һәм язу кунекмәләре" (1993).

During the development of an algorithmic model we had to face a number of difficulties. One of them is the discrepancy between the number of vowels and consonants in two alphabets. There are 9 vowels in the extended Latin alphabet: **a, ä, ü, u, o, i, e, ı, ö**, whereas there are 13 vowels in Tatar language if Cyrillic graphics is used: **а, ә, у, ү, о, ө, ы, е, э, и, я, ю, е**. It is important to note that the letters «я», «ю», «е» are the combination of two sounds: я - [йа], ю - [йу], е - [йы], [йе]. That is, in these cases the law of vowel harmony is violated.

26 letters are used in extended Latin to indicate consonants. Table 1 shows the correspondence between the consonant letters of two alphabets.

Table 1. Compliance with consonants, used in Tatar language (Latin, Cyrillic)

Latin	Cyrillic	Latin	Cyrillic
Bb	Бб	Nn	Нн
Cc	Җҗ	Ņņ	Ңң
Çç	Чч	Pp	Пп
Dd	Дд	Rr	Рр
Ff	Фф	Ss	Сс
Gg	Гг	Şş	Шш
Ğğ	Гъ	Tt	Тт
Hh	Һһ	Vv	Вв
Jj	Ж ж	Ww	Уау
Kk	Кк	Xx	Хх
Qq	Къ	Yy	Йй
Ll	Лл	Zz	Зз
Mm	Мм	Şç	Щщ

Like any living language, Tatar language is being developed. Borrowings and neologisms are an integral part of a language functioning and historical change, one of the main ways to replenish its vocabulary. During the development of a basic conversion principle it is decided to adhere to the principle, based on etymology. In order to determine the origin of Tatar words the following algorithms were chosen: bigram analysis, combined analysis and search. It is proposed to apply different transliteration rules depending on a word origin. Original Tatar words are converted according to the phonetic principle, and borrowed words are converted according to the rules of transliteration.

A specific set of rules is applied to original Tatar words, based primarily on the phonetic principle (you write the sounds you hear) - tavyk - tawıq.

A simplified set of rules, close to the mechanical transliteration (strict conformity between a symbol in Cyrillic and a symbol in Latin) is used for borrowed words (of Arab-Persian, Russian and English origin).

3. RESULTS

Such algorithms of a word origin determination as bigram analysis, combined analysis including bigram and morphemic analysis and also sorting method (brute-force) are selected and modified taking into account the peculiarities of Tatar language system. Let us discuss the advantages and the disadvantages of selected algorithms.

1) Bigram analysis. Bigram is a pair of linguistic units in a row (in our case - a couple of letters). Some bigrams are more common in a language than others, therefore it is possible to determine the origin of a word using statistical methods and analyzing its bigrams. Such algorithms are widely used in web industry [Grechnikov E.A., Gusev G.G., A.A. Kustarev, Raigorodsky A.M., 2001; Attenberg, J., Suel, T., 2008; Benczur, A., Biro, I., Csalogany, K., and Sarlos, T., 2007]. In particular, they are used in Internet browsers to determine a web page coding. The Internet company Yandex uses similar methods to determine the automatically generated text on web pages. A special program was developed to identify bigrams, which analyzed the texts, composed exclusively of native Tatar words.

Of course, this algorithm has its drawbacks. This is due to the fact that Tatar language is an agglutinative one, that is, the dominant principle of word formation is agglutination - the "sticking" of the new morphemes to a word end. Such a principle of word formation creates difficulties at the determination of a word origin.

Example: Let's consider the word "Андрейныкыларгадыр". This word is the borrowed one, and most bigrams are met rarely in a root ("Андрей"). However, a large part of the word is the suffix which comes after the root ("-ныкыларгадыр"). Bigrams presented in suffixed morphemes are used very often in Tatar language. During the use of abovementioned bigram analysis algorithm, the program will identify this word as a native one. Therefore, it is necessary to determine the root of a word, and then analyze only a root.

2) A combined analysis (bigram + morphemic). The principle of this algorithm operation is that a word is divided into a root and a suffix before testing. The suffix is Latinized according to the native language rules,

and the root portion is checked according to bigrams. To isolate a root a special software is used - morphemic analyzer.

3) Search method or Brute-force. Brute-force is the reception in programming at which an original word is checked for correspondence from a preliminary prepared list. When this algorithm is used, you need to create a set from native Tatar words, each word is checked on introduction into the compiled list.

The disadvantage of the algorithm is that it will work only if a word is included in the set. If it is not in the list, the program will determine the initial word as a foreign one. The use of this algorithm leads to the increase of word conversion period.

In the future, during the improvement of a generalized algorithm it is necessary to focus also on such factors as speed and accuracy of an algorithm to determine the origin of a word. In our opinion, the best algorithm among the proposed ones at the moment is the combined method, as it allows you to convert words most accurately without losing the program operation speed.

An algorithmic model of the Tatar word translation from the Cyrillic to the Latin alphabet is developed (Figure 1). This algorithmic model became the basis for the computer program development.

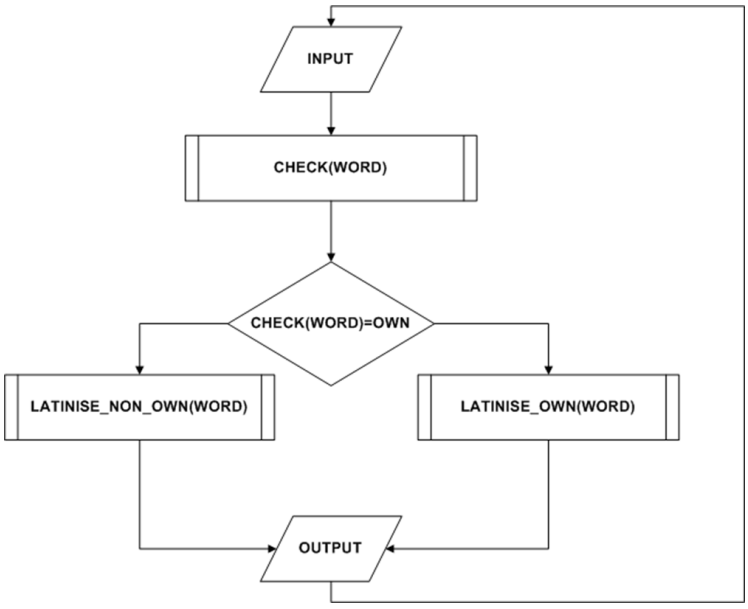

Fig. 1. *Message translation algorithm recorded in Tatar language into Latin using Cyrillic symbols.*

The algorithm is operated as follows:

Input - a program user enters a word in Tatar language, written in Cyrillic.

Check - word etymology check (a native or a borrowed one).

Check (word) = own - conditional selection block. Depending on test results, the word is converted according to the set of rules:

1. case **(Latinise_Own)** - if a word is a native Tatar one, it is converted character by character. This block contains a set of procedures and functions for conversion. All procedures are tailored taking into account the translation rules from the Cyrillic to the Latin alphabet.

2. case **(Latinise_non_Own)** - if a word is a borrowed one, it is converted character by character according to the borrowed word conversion rules. This block contains a set of procedures and functions which realize the process of mechanical transliteration.

Note! Printed characters, but not letters (punctuation marks, numbers) are not processed.

Output block - a message is displayed written in Latin alphabet.

Then a next word is read (return to **Input** block).

The developed algorithmic model and the principles formed the basis for the creation of a computer program. A group was created for its development in 2015, consisting of educational technology department members and information systems in philology and the students and masters of the Institute of Philology and Intercultural Communication at Kazan Federal University. The group includes the programmers and specialists in the field of Tatar linguistics. The application is implemented using an integrated development environment (IDE) Embarcadero Delphi 2009. At the moment, we have developed a software product that allows you to convert native Tatar words into Latin. An example of the program is presented on Figure 2.

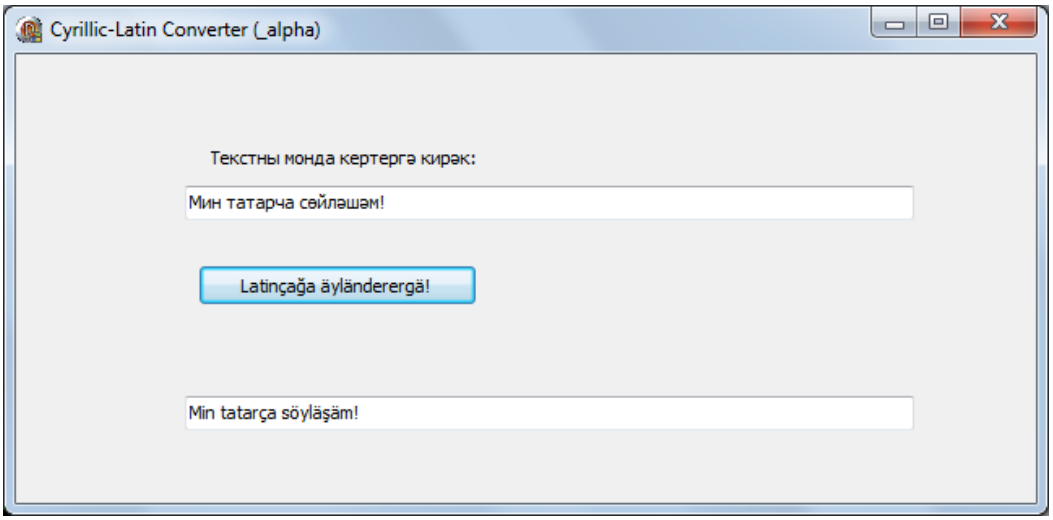

Fig. 2. Converter operation demonstration

4. CONCLUSIONS

At the moment, the problem of Tatar language word etymology determination is not solved completely. A more sophisticated checking algorithm will be developed. A program product is developed which allows you to carry out the conversion process. In addition to the software for the Windows operating system, it is planned to develop the mobile applications for mobile operating systems (iOS, Android or Windows Phone / Mobile), an Internet resource which provides the ability to convert in on-line mode, and the add-ins for common text processor Microsoft Office Word.

5. SUMMARY

The development of this software will open new opportunities for Tatar language use in the field of written communications, IT industry and education. The use of the developed software, along with other solutions [Danilov, A., Salekhova, L., 2015; Zaripova, R., Salekhova, L., Tuktamyshov, N., Salakhov, R. 2014; Nevzorova, O., Suleymanov, D., Gilmullin, R., Gatiatullin, A., Khakimov, B., 2013; Fatkhullova K., Zamaletdinov R., Yusupova A., 2013], will improve the level of information culture among Tatar-speaking Internet users and information technologies, the transliteration of the Tatar language texts will enable the domestic foreign scientists to conduct joint research together.

ACKNOWLEDGEMENTS

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

REFERENCES

1. Hakov V.H. Теленбелгэн ил ачар: Латин графикасында уку һәм язу күнекмәләре [Text]/ Hakov V.I. – Kazn: Megarif publishing house, 1993 – 140 p.
2. Information materials on the final results of the National Population Census in 2010 / Website of the Federal State Statistics Service, 2010, URL: http://www.gks.ru/free_doc/new_site/perepis2010/perepis_itogi1612.htm date of appeal: 22.01.2015)
3. About the use of the Tatar language as a state language of the Republic of Tatarstan [electronic resource]: the law of the Republic of Tatarstan issued on January 12, 2013. №1-RTL - Access mode: http://mon.tatarstan.ru/rus/file/pub/pub_227812.pdf
4. Nevzorova, O., Suleymanov, D., Gilmullin, R., Gatiatullin, A., Khakimov, B. Tatar National Corpus “Tugantel”: structure and features of grammatical mark-up // Procedia - Social and Behavioral Sciences. Vol. 95. Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013) // Ed. Chelo Vargas-Sierra, pp. 68-74.
5. Danilov, A., Salekhova, L. Design of Virtual Keyboard for Tatar-Speaking Users on the Basis of the Mobile Operating System Android// International Journal of Soft Computing, №10 (5): p.348-352, 2015.

6. Grechnikov E.A., Gusev G.G., A.A. Kustarev., Raigorodsky A.M. Search of unnatural texts // Proceedings of VLDB-2001, 2001, 306-308.
7. Attenberg, J., Suel, T. Cleaning search results using term distance features // Proceedings of AIRWeb-2008, p. 21-24.
8. Benczur, A., Biro, I., Csalogany, K., and Sarlos, T.Web spam detection via commercial intent analysis. //Proceedings of AIRWeb-2007, NewYork, USA, 2007, p. 89-92.
9. Zaripova, R., Salekhova, L., Tuktamyshov, N., Salakhov, R. 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/ (date of appeal: 23.03.2016)
10. Fatkhullova K.S., Zamaletdinov R.R., Yusupova A.S. Information-Communicative Devices for Tatar Language Teaching // World Applied Sciences Journal, 2013, Volume 26, Issue 1, pp. 103-107.