

# Formal Models and Software Tools for the Computer Processing of the Tatar Language

**Suleymanov D., Khusainov A., Gilmullin R.**

Institute of Applied Semiotics of the Tatarstan Academy of Sciences,

Kazan Federal University

Kazan, Russia

{dvd.t.slt, khusainov.aidar, rinatgilmullin}@gmail.com

## Abstract

The paper describes models and software tools developed mainly as part of the State program for the preservation, study and development of the state languages of the Republic of Tatarstan and other languages in the Republic of Tatarstan. The content of the paper shows the current state of work on creating software tools and systems to support the Tatar language in computer technologies. The introduction provides a brief chronological summary of the results of research and development of the Institute of Applied Semiotics of the Tatarstan Academy of Sciences on the national computer systems localization and the use of the Tatar language in Information Technology, showing their compliance with modern trends in the development of natural language processing technology.

**Keywords:** Tatar language, language resources, NLP tools

## Абстракт

Мәкаләдә Татарстан Республикасы дәүләт телләрен һәм Татарстан Республикасындагы башка телләргә саклау, өйрәнү һәм үстерү буенча Татарстан Республикасы дәүләт программасын гамәлгә ашыру кысаларында эшләнгән модельләр һәм программалар тасвирлана. Мәкаләнең эчтәлегенә компьютер технологияләрен татар теле белән тәмин итү өчен төзелгән программалар һәм системаларның бүгенге торышын чагылдыра. Керештә Татарстан Республикасы Фәннәр академиясе гамәли семиотика институтының компьютерларны татарчалаштыру һәм татар телен инфокоммуникация технологияләрендә куллану юнәлешендәге тикшеренүләре һәм эшләнмәләре турында аңлатма бирелә, аларның заманча технологияләр нигезендә башкарылуы күрсәтелә.

## 1. Introduction

Natural Language Processing (NLP) is the general direction of artificial intelligence and mathematical linguistics. Currently, for the most spoken languages of the world, specialized software tools have been developed to support national languages in information technology.

On the other hand, there is a class of low-resourced languages that suffer from a lack of available language resources and software. Of particular interest to the Republic of Tatarstan are technologies that support the Turkic languages, including the Tatar language that is also known as a low-resourced language.

The work in the field of Tatar NLP began in early 1990s in Tatarstan Academy of Sciences and Kazan State University. In this paper we present the main results obtained during this time, as well as basic information about the Tatar language and the short summary of the development history.

### 1.1 The Tatar Language

Tatar is the second spoken language in Russia. There are 4.2 million of speakers in Russia and near 5 million of speakers in the world (Eberhard et al., 2019). The Cyrillic alphabet (unified in 1939) consists of 39 characters. There are 12 vowel and 28 consonant sounds. Different dialects of Tatar can be identified: Western, Kazan (Middle) and Eastern. Based on the existing language classification (Berment, 2004; Krauwer, 2003), it was assigned to the under-resourced language class (Khusainov, 2014).

However, recent results in machine translation, speech analysis and synthesis fields can change this situation.

### 1.2 History

Research and development in the field of computer linguistics for Tatar began in 1993 as part of the Joint Laboratory for Artificial Intelligence of the Academy of Sciences of the Tatarstan Republic and Kazan State University, which in 2009 was transformed into the Institute of Applied Semiotics of the Academy of Sciences of the Republic of Tatarstan.

All these years, one of the most important scientific and applied problems has been the development of software tools and linguistic resources for the widespread use of the Tatar language in information technology, including mobile devices and the Internet. The particular importance of the Tatar localization of computer technology is also determined by the need to ensure the parity functioning of the Tatar language along with the Russian language as the state language in the Republic of Tatarstan.

Research and development in computer technology for the Tatar language began in the end of the 1980s from the implementing of the first monitor and printer drivers, a text editor and the Tatar spellchecker, localization of computer publishing systems needed for the publication of Tatar books, newspapers and magazines.

It is interesting to look at the chronology of early research and development:

– 1992 year. Multimedia compact discs with the Tatar language teaching program “My First Tatar Dictionary”,

as well as the educational multimedia CD “Tatar Telle Zaman”, which includes, in addition to the training, testing, and dialogue blocks, also a block of linguistic games. All implementations provide training in trilingual (Tatar, Russian, English) mode. Moreover, the Tatar language operates in two modes: on the basis of the Cyrillic alphabet and on the basis of the Latin alphabet.

– 1994 year. The concept and architecture of the machine fund of the Tatar language were developed and the filling of the machine fund with electronic dictionaries and texts, with modules for processing Tatar texts, data processing systems in the Tatar language was started.

– 1995 year. A spellchecker for Tatar texts has been developed, which allows to find and correct spelling errors in Tatar texts.

– 1996 year. Joint with ABBYY company, a version of the FineReader OCR program was developed for Tatar.

– 1998 year. Joint with the Belkent University (Turkey), a Tatar two-level analyzer was developed that performs morphological analysis and generation of Tatar word forms.

– 2000 year. A diphone-based Tatar speech synthesizer was developed.

– 2003 year. A complete structural and functional model of the Tatar morphology was created.

## 2. Tatar Localization

National localization of computer systems is very important for the preservation and development of low-resource languages, as well as a necessary base for the development of language resources, natural language processors and programs for e-learning.

The most important and knowledge-intensive step in localization is the development and adoption of standards for knowledge representation and of a terminology for computer science and information technology.

The main criteria for Tatar localization are correctness, accuracy of the translation of the text, its semantic correspondence to the original text; brevity and clarity of texts (instructions, actions) on the menu buttons; clarity and compactness of texts in reference files. Typical derivational and syntactic models were developed for the different screen components, and the overall style of the interface was created on this basis.

In order to create and use linguistic resources for low-resource languages in all spheres of their manifestation (such as science, education, publishing, information processing etc.), it is important to integrate knowledge and skills of experts in related spheres: computer science, mathematics and linguistics.

Currently, the following important linguistic resources and software have been developed and are actively used while being under further development. Thanks to the cooperation of the Academy of Sciences of the Tatarstan Republic with Microsoft company, all versions of the OS Windows, beginning from the Windows NT, were localized.

The Tatar language became the second Turkic language after the Turkish language, localized by specialists of the

republic itself, and not by the developers of the company. Microsoft Office applications were also localized, including the interface and help files, as well as the Tatar spellchecker.

Currently, the Tatar language is being actively introduced into mobile devices. Localized service applications are developed: keyboards, dictionaries, predictive typing systems, games, tutorials. In 2016, Tatar localization of the Russian mobile operating system Aurora OS began (jointly with the “Open Mobile Platform” company). Aurora OS has become the first mobile operating system that makes it possible to fully use the Tatar language along with Russian in mobile devices.

Another unique software system adapted for the Tatar language by our specialists together with «ABBYY LS» company is SmartCAT professional translation system (<http://smartcat.ai/>). The system is intended for widespread use as a tool for a professional translator with various useful functions (machine translation, electronic dictionaries, etc.). This platform for automation of translation, which optimizes the work process and comprehensively solves all translation tasks, allowing you to create projects, monitor the work of the translation team in real time, check translated segments, discuss details with the team directly in the system. The SmartCAT system is currently implemented in organizations and departments in all regions of the Republic of Tatarstan.

## 3. Software and Linguistic Resources for Tatar

The development of software tools, applications, and linguistic resources for the Tatar language ensures the use of computer systems and technologies for working with the Tatar language in all spheres and forms of its manifestation.

### 3.1 Tatar National Corpus “Tugan Tel”

The Tatar corpus “Tugan Tel” is a linguistic resource of the modern literary Tatar language. The corpus is addressed to a wide range of users: linguists, specialists in the field of Tatar, Turkic and general linguistics, typologists, teachers of the Tatar language, as well as everyone who studies and is interested in the Tatar language.

The volume of the corpus is over 200 million word forms and contains texts of various genres (fiction, media texts, texts of official documents, textbooks, scientific publications, etc.). Each document has a meta description (authors, their gender, creation dates, genres, parts, chapters, etc.). The texts included in the corpus are provided with morphological annotation (information about the part of speech and the grammatical characteristics of the word form). Morphological annotation of corpus texts is carried out automatically using the module of two-level morphological analysis of the Tatar language, implemented via PC-KIMMO software toolkit.

A search system has been developed for the corpus that allows searching for material by lexeme, word form, as well as by individual grammatical characteristics. Corpus is available at <https://tugantel.tatar>.

### 3.2 Electronic Version of the Atlas of Tatar Dialects

In 2011-2012, an electronic version of the atlas of Tatar dialects was created. The Atlas includes all the main areas of Tatars settlement and reflects information on the phonetics, morphology, vocabulary and syntax of the Tatar language, collected in 28 regions of Russia.

The electronic Atlas database contains information on the distribution of the meanings of 215 linguistic phenomena across 1047 settlements. Maps display the features of Tatar dialects in the phonetics (68 cards), morphology (49 cards), vocabulary (93 cards) and syntax (5 cards) sections. The maps of the Atlas provide information on the distribution of dialects in the selected settlements.

The release of the electronic version of the atlas is a new stage in the presentation of dialectical knowledge of the Tatar language based on geographic information systems. An electronic Atlas is available at <http://atlas.antat.ru/>.

### 3.3 Morphological analysis and disambiguation

The Tatar language has a rich and regular, almost automatic, morphology (Suleymanov, 1998). The morphological model of the Tatar language is a basic component in almost all fully functional linguistic analyzers. Accordingly, the creation of a computer model of the Tatar morphology was one of the first and important tasks. Given the structural specificity of the Tatar language and based on applied problems, three different morphological models have been developed to date.

Generative morphology model based on affixing rules, although inferior to other models in speed, provides the completeness of the analysis of the word form, allowing you to fully take into account the agglutinative nature of the language, recognizing word forms of potentially unlimited length.

The paradigmatic model of Tatar morphology provides quick recognition of word forms and analysis of the correctness of Tatar word forms with an accuracy of 95% and is used in MS Windows and its Office applications. In addition, in a joint project with the Belkent University (Turkey), a two-level model of the Tatar language morphology was developed, implemented via the PC KIMMO software shell. A hybrid model of morphological analysis has also been created, using generative and paradigmatic approaches, which is part of the "Tatar morpheme" information system.

### 3.4 Tatar Speech Recognition and Synthesis

Systems of automatic recognition of continuous speech and its synthesis allow to carry out work on the implementation of human-machine speech interface.

A set of speech technologies is being developed at the Institute of Applied Semiotics, which includes the ability to identify the language of the speaker, automatic speech

recognition and synthesis for Tatar. Databases of textual and speech information in the Tatar language are accumulated and analyzed, machine learning technologies are developed, the speech interface in the Tatar language is integrated into modern PCs and mobile devices.

Further development of speech technologies will open up prospects for sharing research results in the field of semantic analysis of text in the Tatar language, and will allow the creation of intelligent systems for helping the visually impaired, speech translators, intellectual assistants, etc.

### 3.5 Russian-Tatar Neural Machine Translation

Russian and Tatar languages are the official languages in the Republic of Tatarstan. This fact makes urgent the task of providing the population, state and other institutions with the possibility of automatic translation between these languages.

One of the key areas of activity of the Institute of Applied Semiotics of the Tatarstan Academy of Sciences is the creation of a machine translator in a Russian-Tatar language pair. An important component of the machine translator is linguistic support, which currently includes 1 million Russian-Tatar pairs. The created version of the Russian-Tatar translator (<https://translate.tatar>) is currently the best among the analogues in terms of translation quality.

The results of constructing a machine translator system for the Russian-Tatar language pair show that modern neural network algorithms and approaches are able to solve the translation problem at a fairly high level even for low-resourced language pairs.

## 4. Conclusion

The article describes the models and software tools developed for the Tatar language.

The description reflects the current state of work on the development of software tools and systems to support the Tatar language in computer technologies and includes a description of the morphological analyzer of the Tatar language, the system of machine translation for the Russian-Tatar language pair, the synthesizer and recognizer of Tatar speech, the electronic corpus of the Tatar language "Tugan tel", as well as the electronic Atlas of Tatar dialects. We also give an overview of another important part of research and work on the localization of software products: operating systems (including mobile), mobile applications, websites.

## 5. Bibliographical References

- Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2019. *Ethnologue: Languages of the World*. Twenty-second edition. Dallas, Texas: SIL International. Online version: <http://www.ethnologue.com>.
- V. Berment, "Méthodes pour informatiser des langues et des groupes de langues peu dotés", Ph.D. Thesis, J. Fourier University, Grenoble I, 2004.
- S. Krauwer, "The basic language resource kit (BLARK) as the first milestone for the language resources

- roadmap”, In Proc. of International Workshop Speech and Computer SPEECOM, Moscow, Russia, 2003, P. 8–15.
- A. Khusainov, “Tekhnologiya avtomatizatsii sozdaniya I otsenki kachestva programmnikh sredstv analiza rechi s uchetom osobennostey maloresursnykh yazikov”, Ph.D. Thesis, Kazan, 2014, 162 p.
- D. Suleymanov, “Formalnaya elegantnost I estestvennaya slozhnost morfologii tatarskogo yazyka”, In. Proc. of Information Technology in the Humanities, Kazan, Russia, 1998. URL: [http://www.kcn.ru/\\_tat\\_ru/universitet/gum\\_konf/ot7.htm](http://www.kcn.ru/_tat_ru/universitet/gum_konf/ot7.htm).