DOI: 10.15514/ISPRAS-2022-34(6)-13

Research Perspectives on the Tatar language based on the LingvoDoc platform

[1,2] *F.Sh. Nurieva, ORCID: 0000-0001-9957-9734 <fanuzanurieva@yandex.ru>*
[1] *G.R. Galiullina, ORCID: 0000-0001-6923-2190 <caliullina@list.ru>*
[1] *A.F. Yusupov, ORCID: 0000-0003-0363-5303 <faikovich@mail.ru>*

[1] *Kazan Federal University,*
18 Kremlevskaya str., Kazan, 420008, Russian Federation
[2] *Ivannikov Institute for System Programming of the RAS,*
25, Alexander Solzhenitsyn Str., Moscow, 109004, Russia

Abstract. The article discusses research perspectives on the Tatar language based on the LingvoDoc platform. Digitalization of language learning in modern linguistics allows us to move to a new level of describing the language structure. Large corpora containing millions of word forms have been created in all European languages since the 90s of the last century. Currently, this has been done not only in the Russian language, but also in many national languages of Russia such as Tatar, Bashkir, Udmurt, Mari, Moksha, Komi, etc. One of the recognized platforms in modern national linguistics is the development of the LingvoDoc virtual laboratory, created ISP RAS. This platform gives an opportunity to create, store and analyze multilayer dictionaries, language materials and dialects. The main functionality of Lingvodoc is used by more than 250 linguists who process their materials online, more than 1000 dictionaries and 300 text corpora in the national languages of the Russian Federation have already been collected. We consider the possibilities of this platform to study the Tatar language. We believe that electronic corpora allow us to solve a variety of theoretical and practical problems of the language. At present, when the Tatar literary and everyday spoken language is actively used in all fields, it is very important to make a complete description of its features, which will help create more accurate grammars and dictionaries. The relevance of the study is due to the need to use a gloss corpus of texts in the Tatar language. As modern studies in linguistics show, nowadays it is impossible to describe the state of the language without such corpora and analyze its grammatical structure, which corresponds to the world standards of modern science. The LingvoDoc platform makes it possible to process a significant amount of material in a short time and create corpora with glossing and removed homonymy based on samples of the Tatar literary, business, colloquial and dialect languages.

Keywords: Tatar language; LingvoDoc; corpus of the Tatar language; grammar; colloquial speech

For citation: Nurieva F.Sh., Galiullina G.R., Yusupov A.F. Research Perspectives on the Tatar language based on the LingvoDoc platform**.** Trudy ISP RAN/Proc. ISP RAS, vol. 34, issue 6, 2022. pp. 173-178. DOI: 10.15514/ISPRAS-2022-34(6)-13

Перспективы исследований татарского языка на платформе LingvoDoc

[1,2] *Ф.Ш. Нуриева, ORCID: 0000-0001-9957-9734 <fanuzanurieva@yandex.ru>*
[1] *Г.Р. Галиуллина, ORCID: 0000-0001-6923-2190 <caliullina@list.ru>*
[1] *А.Ф. Юсупов, ORCID: 0000-0003-0363-5303 <faikovich@mail.ru>*

[1] *Казанский (Приволжский) федеральный университет*
420008, Россия, г. Казань, ул. Кремлевская, д.18
[2] *Институт системного программирования им. В.П. Иванникова РАН,*
109004, Россия, г. Москва, ул. А. Солженицына, д. 25

Аннотация. В статье рассматриваются перспективы исследования татарского языка на платформе LingvoDoc. Цифровизация изучения языка в современной лингвистике позволяет перейти на новый уровень описания структуры языка. С 90-х годов прошлого века во всех европейских языках созданы большие корпуса, содержащие миллионы словоформ. В настоящее время это сделано не только в русском языке, но и во многих национальных языках России, таких как татарский, башкирский, удмуртский, марийский, мокшанский, коми и др. Одной из признанных площадок в современном отечественном языкознании является разработанная в ИСП РАН виртуальная лаборатория. Эта платформа дает возможность создавать, хранить и анализировать многослойные словари, языковые материалы и диалекты. Основным функционалом LingvoDoc пользуются более 250 лингвистов, обрабатывающих свои материалы онлайн, уже собрано более 1000 словарей и 300 корпусов текстов на национальных языках РФ. Мы рассматриваем возможности этой платформы для изучения татарского языка. Мы считаем, что электронные корпуса позволяют решать самые разные теоретические и практические проблемы языка. В настоящее время, когда татарский литературно-бытовой разговорный язык активно используется во всех сферах, очень важно сделать полное описание его особенностей, что поможет создать более точные грамматики и словари. Актуальность исследования обусловлена необходимостью использования глоссового корпуса текстов на татарском языке. Как показывают современные исследования в области языкознания, в настоящее время невозможно описать состояние языка без таких корпусов и проанализировать его грамматический строй, соответствующий мировым стандартам современной науки. Платформа LingvoDoc позволяет в сжатые сроки обрабатывать значительный объем материала и создавать корпуса с глоссированием и снятием омонимии на основе образцов татарского литературного, делового, разговорного и диалектного языков.

Ключевые слова: татарский язык; LingvoDoc; корпус татарского языка; грамматика; разговорная речь

Для цитирования: Нуриева Ф.Ш., Галиуллина Г.Р., Юсупов А.Ф. Перспективы исследований татарского языка на платформе LingvoDoc. Труды ИСП РАН, том 34, вып. 6, 2022 г., стр. 173-178. DOI: 10.15514/ISPRAS-2022-34(6)-13

1. Introduction

Modern linguistic science pays more attention to the language research using digital technologies. Such technologies make it possible to model the “functioning of a language in certain conditions” [1]. As is known, computational linguistics has various developing areas, such as automatic analysis and synthesis of texts, creating and maintaining electronic dictionaries, creating linguistic databases and electronic corpora of languages, etc., they all aimed at a comprehensive analysis of linguistic phenomena [2; 3], as well as the preservation of linguistic facts.

Language corpora are considered to be one of the forms of preservation and ordering of linguistic facts. Electronic corpora along with linguistic information make it possible to solve a variety of theoretical and practical problems of the language. This area is becoming one of leading in modern linguistic research. Our research considers the achievements of national science in this area, but we see that there is the lack of the most complete corpus of the Tatar language with its dialects and colloquial speech, so our study based on LingvoDoc will fill the existing gap in here [4].

Russian linguistics currently has corpora for the Russian language (cf. https://ruscorpora.ru/new/), as well as for many national languages of Russia:

- Tatar (cf. http://www.tugantel.tatar/, https://www.corpus.tatar/),
- Bashkir (cf. https://bashcorpus.ru/),
- Udmurt (cf. http://udmcorpus.udman.ru/, http://udmurt.web-corpora. net/),
- Mari (cf. http://corp.marnii.ru/, http://meadow-mari.web-corpora.net/),
- Erzya (http://erzya.web-corpora.net/) Moksha (http://moksha.web-corpora.net/),
- Komi (cf. http://komi-zyrian.web-corpora.net/)

and corpus work is being actively conducted in other national languages.

Ivannikov Institute for System Programming of the Russian Academy of Sciences (ISP RAS) developed a platform for digital data processing in linguistics called LingvoDoc. At present, the LingvoDoc platform (lingvodoc.ispras.ru) allows users to upload a document in Word format and process it using a parser. The available text corpora such as http://www.tugantel.tatar/, https://www.corpus.tatar/ are a linguistic resource of the modern literary Tatar language, but they don't provide complete morphological tagging. Creating corpora of fiction, official texts, dialects and colloquial speech with glossing and removed homonymy using LingvoDoc will allow us to evaluate existing dictionaries and collections of texts in terms of their reliability. Creating electronic corpora, along with linguistic information, is becoming one of the main methods of modern linguistic research.

2. State of art in Tatar language research and perspectives of LingvoDoc use

At present, when the Tatar literary and everyday spoken language is actively used in all fields, it is very important to make a complete description of its features, which will help create more accurate grammars and dictionaries. The relevance of the study is due to the need to use a gloss corpus of texts in the Tatar language. As modern studies in linguistics show, nowadays it is impossible to describe the state of the language without such corpora and analyze its grammatical structure, which corresponds to the world standards of modern science. The LingvoDoc platform makes it possible to process a significant amount of material in a short time and create a corpora with glossing and removed homonymy based on samples of the Tatar literary, business, colloquial and dialect languages.

Our study analyzes the achievements of national science in this field. Research based on LingvoDoc will fill the existing gap that appeared due to the lack of the most complete corpus of the Tatar language with its dialects and colloquial speech.

2.1 Grammar-oriented research

Tatar language grammar and its dialects are fully studied well enough. In recent years, research has been carried out on topic-comment articulation, the semantic structure of the sentence, the syntax and text style as the main unit of speech. The results of grammatical studies are shown in various types of descriptions. Traditionally, they are distinguished as scientific, descriptive and normative.

Scientific grammar includes historical grammar, studying the structure of the language in development or at its individual stages in the past. This field was led by L. Zalay, V.Kh. Khakov, F.S. Faseev, I.A. Abdullin, D.G. Tumasheva, F.M. Khisamova, F.A. Ganiev and others.

Comparative (contrastive) grammar is also a part of scientific grammar. Tatar language grammar describes the similarities and differences in Tatar and other languages, rather than in Russian (grammar books by K.Z. Zinnatullina, E.M. Akhunzyanova, L.K. Bayramova, etc.). The academic book "Tatar Grammar " was published in 3 volumes in 1992-1993 (State Prize of the Republic of Tatarstan, 1994). It contains a description of the phonetic and grammatical structure of the modern Tatar literary language. The Ibragimov Institute of Language, Literature and Art published "Academic lexicology" in 3 parts (2015–2018), "Academic Grammar" in 3 parts (2015–2017) and others, but there are quite a lot of unresolved issues.

2.2 Tatar language morphology

Problems in Tatar language morphology, in particular, the criteria and principles for identifying parts of speech, the interaction of units of different classes, transitional phenomena in the parts of speech system can be solved using the LingvoDoc toolkit. Modern technical means make it possible to present texts in various formats, and the current level of linguistic knowledge enables us to classify and index linguistic data, which allows us to present the collected material in a complex way, i.e. in the form of multimedia marked-up corpora and dictionary databases. The language features identified in this way demonstrate the system and structural, communicative and functional characteristics of the language.

An urgent task for Tatar morphology is the study of parts of speech as a continuum consisting of core and peripheral elements, which combine different word orders based, first of all, on their functional significance. The research will provide an opportunity to identify the process of diffusion and mutual transition between classes of words of certain speech parts. In the Tatar language, the problem of identifying a separate group of words is still controversial, for example, words like altyn 'gold' – altyn baldak 'golden ring', tash 'stone' – tash kuper 'stone bridge', agach 'tree' – agach өy 'wooden house', etc. Thus, the study using the LingvoDoc toolkit will allow us to solve not only the most confusing problems of the modern theory of part-of-speech attribution of words in the Tatar language, but also it will give an opportunity to present our view on solving such problems.

2.3 Colloquial speech

Created electronic corpora of Tatar fiction, colloquial speech and official business texts with full gloss and removed homonymy, as well as audio dictionaries of spontaneous colloquial speech with transcription in IPA, audio recordings of contexts will make it possible to determine the conceptual and semantic potential and functional features of language units, both in individual subsystems of the language, and in the language as a whole. So, for example, data on the functioning and frequency of individual classes of words and lexemes that function in different subsystems of the language allow solving some controversial issues in the theory of parts of speech, such as mutual transition 'noun-adjective', 'adjective-adverb', modal words, and onomatopoeic words in the Tatar language. These results will clarify theoretical issues in the field of grammar and word formation of the Tatar language.

Research models of the vocabulary system of the language are designed to model and reflect the real relationship between the functional-semantic types of words in all the variety of connections between them. Research on the LingvoDoc platform opens up opportunities to combine all elements: nominative, connective, etc. because these units have a functional value.

One of the promising tasks is to present a separate corpus of texts of Tatar spontaneous colloquial speech, characteristic of different social and age groups living in rural and urban areas.

As is known, the national language manifests itself in literary forms, in territorial and social dialects, and in vernacular, they all interact and meet various communicative needs of society. Modern Tatar linguistics practically doesn't have any comprehensive studies of the spoken language based on these territorial and social dialects and colloquial forms. At the present stage, the influence of dialects and forms of the spoken language in the literary language and its norms has sharply increased in the Tatar language. As part of the project, we intend to identify the main forms and features of the spoken language (phonetic, lexical, semantic, derivational, and grammatical) and study their implementation in the communicative process of individual social groups.

"The need to study Tatar colloquial speech with all its specific subsystems, structural and content diversity is due to several factors. Let's look at some of them. In modern society, there is a tendency to spread Tatar colloquial speech at home, in the media and electronic communication. There is a noticeable tendency for elements of an uncodified language to penetrate into the literary language. In the context of the integration of languages and dialects, there is a need for a systematic and comprehensive study of the Tatar spoken language in order to reveal the specifics of modern Tatar

speech, determine the boundaries of the territorial and social dialects of the Tatar language, identify the mechanisms of mutual influence and diffusion of the speech characteristics of the ethnicities living in the Republic of Tatarstan" [5: 5]. Spoken language will be presented in transcription and audio or video dubbing, which will give the oral corpus an accurate reproduction format. The inclusion of oral speech in the form of corpus texts will open up opportunities for developing theoretical problems of the current state of the Tatar language and determining the prospects for development in the context of globalism. All results will be available on the LingvoDoc platform.

It is known, the norm of the codified language acts as a regulator of the interaction between literary forms of the language and dialects. However, the norm itself is subject to changes over time and depends on the nature of the processes that take place in society. Society development is not only reflected in the language, but also has a significant impact on the further evaluation of its forms and elements. Unfortunately, traditional studies of the grammatical and lexical structure are behind the changes in modern colloquial speech. Because the volume and specificity of the material presented in this kind of research may not be full-scale.

3. Conclusion

Electronic corpora make it possible to fully reflect the language processes in different subsystems of the language. Thus, we will be able to predict in time the further development of the structure and possible changes in the norms of the codified Tatar language.

References / Список литературы

[1] Baranov A.N. Introduction to Applied Linguistics: Learning Guide. 3rd ed. Moscow, LKI, 2007, 360 p. (in Russian) / Баранов А.Н. Введение в прикладную лингвистику: учебное пособие. 3-е изд. М., ЛКИ, 2007 г., 360 стр. /

[2] Zubov A.V., Zubova I.I. Information technology in linguistics: Learning Guide. Moscow, Academy, 2004, 208 p. (in Russian) / Зубов А.В., Зубова И.И. Информационные технологии в лингвистике: учебное пособие. М., Академия, 2004 г., 208 стр.

[3] Bolshakov I A., Gelbukh A. Computational Linguistics. Models, Resources, Applications. Mexico, IPN-UNAM-FCE, 2004, 186 pp.

[4] Normanskaja J.V. Frst turkic cyrillic books on the LingvoDoc platform. Native languages and Cultures in the modern changing world, issue 1, 2022, pp. 43-57 / Норманская Ю.В. Первые тюркские кириллические книги на платформе ЛингвоДок. Родные языки и культуры в современном изменяющемся мир, вып. 1, 2022 г., стр. 43-57.

[5] Galiullina G.R., Kadirova E.Kh., Khadieva G.K. Modern Tatar colloquial speech: identification features and social differentiation. Kazan, Kazan University Publishing House, 2022, 222 p. (in Russian) / Галиуллина Г.Р., Кадирова Э.Х., Хадиева Г.К. Современная татарская разговорная речь: идентификационные признаки и социальная дифференциация. Казань, изд-во Казанског университета, 2020 г,, 222 стр.

Information about authors / Информация об авторах

Fanuza Shakurovna NURIEVA, Doctor of Philology, Professor, Chief Researcher of ISP RAS, Professor of Kazan Federal University. Research interests: Tatar language, turkology, dialectology, comparative and historical linguistics

Фануза Шакуровна НУРИЕВА, доктор филологических наук, профессор, главный научный сотрудник ИСП РАН, профессор КФУ. Научные интересы: татарский язык, тюркология, диалектология, сравнительно-историческое языкознание

Gulshat Raisovna GALIULLINA, Doctor of Philology, Professor, Head of the Department. Research interests: Tatar language, languages of the peoples of the Russian Federation, lexicology, semasiology, onomastics, linguoculturology, sociolinguistics, ethnic culture

Гульшат Раисовна ГАЛИУЛЛИНА, доктор филологических наук, профессор, заведующий кафедрой. Область научных интересов: татарский язык, языки народов РФ, лексикология, семасиология, ономастика, лингвокультурология, социолингвистика, этническая культура.

Airat Faikovich YUSUPOV, Doctor of Philology, Associate Professor. Research interests: Philology, literary criticism, textual criticism, semiotics, Islamic studies.

Айрат Фаикович ЮСУПОВ, доктор филологических наук, доцент. Сфера научных интересов: филология, литературоведение, текстология, семиотика, исламоведение.