

PRE-CORPUS PROCESSING OF THE CRIMEAN TATAR TEXTS

Lemara Selendili,

Sholokhov Moscow State University of Liberal Arts,
16-18 Verkhnyaya Radischevskaya Str., Moscow, 109240, Russia,
lemara2002@hotmail.com.

The article is devoted to the development of the syntactic constructs theory on the material of the Crimean Tatar language. Pre-corpus processing of the linguistic material is done with the aim of creating a theoretical basis for designing appropriate systems of machine translation from/into Crimean Tatar, academic electronic dictionaries and instrumental systems aimed at developing the language in the practice of teaching languages using modern information technologies. In the process of designing models on the basis of huge data, the syntactic constructs' requirements, conditions and sequence of automatic/semitautomatic syntactic marking of the text are taken into consideration. The environment for the realization of the conception of syntactic constructs is the module "Syntax and Phraseology" created for the database of "Russian-Crimean Tatar Dictionary of Linguistic Equivalents".

Key words: the Crimean Tatar language, syntax, structural, applied and mathematical linguistics, linguistic corpus, pre-corpus text processing.

Syntax, being one of the fundamental linguistic sciences, is an inseparable part of the theory of any language. It is a field of paramount importance for languages which have fairly recently become the subject of theoretical studies. Among such languages we should point out Crimean Tatar; as far as we know, the syntax of this language has been studied scientifically only in the dissertation of E.S.Akmollayev [1] (apart from our papers).

The creation of the syntactic theory of constructs in Crimean Tatar sentences is gaining a specific meaning in the view of the modern universal theory "Linguistics of constructs" of E.V.Rakhilina. It is aimed at studying and describing real natural languages in all their aspects and cut-offs on all levels and is designed "*for the following kinds of practice: vocabulary, grammar, corpora studies, expedition and typological research*" [2: 74]. It plays an important role for applied linguistics, which is based on the formalizing of the data on language objects: the plane of expression and plane of content of syntactic units, their morphological, semantic and syntactic peculiarities, formulas and specific combinations of signs (symbols). Such an approach requires a precision of description "as soon as the construction is a complex linguistic object in the formation of which all its components may participate, even if they constitute different language levels" [2: 538]. "*To do this, it is, by all means, better to rely on a construct, not on natural word usage, which is far from being formal*" [2: 238].

Linguistic theory is actively used in the creation of computer didactic programs in different languages not only native for the users, but also

foreign, in developing various lexicographic systems, corpora and means of automatic processing of natural languages, which provide the exchange of information and communication between a computer and a human being in the natural language.

The technologies of corpus linguistics are developing rapidly these days. The scholars raise the problems of the automatic semantic and syntactic layout of the linguistic corpus texts trying to find ways of settling them. Besides, to create the programs of automatic translation, corpus linguistic studies, based on parallel texts, are being conducted more and more frequently and multilingual dictionaries and grammars of different languages are also being implemented.

It is noteworthy that not all Turkic languages have academic grammars and modern dictionaries (electronic, not even paper sometimes), and that is why some of them need pre-corpus processing.

Regarding this, we can say that lexicography and syntax of Crimean Tatar have merely been studied in fragments. A.M.Emirova points out that "*the dictionaries of the present time vividly illustrate the miserable conditions of Crimean Tatar linguistics*" and that "*only based on representative dictionaries, which reflect all the parameters of the language system, objective scientific description and study of any language is possible*" [3]. There are no modern academic dictionaries, thesaurus, ideographic and explanatory dictionaries in the Crimean Tatar language [4]; the syntactic structure of the language has not been studied in detail, no research has been done on the functioning of word combinations and the place of speech patterns and

vocatives are yet to be established alongside the issue of the Crimean Tatar sentence parts; “such parts of a language system as phonetics, phonology, lexicology, phraseology, grammar, lexicography, etc. are in need of thorough study with modern conceptual terminological apparatus” [3].

According to the UN data of December 2010 with reference to the UNESCO interactive atlas “World Endangered Languages”, the Crimean Tatar language is under threat of extinction, especially its Nogai dialect (the northern steppe dialect of Crimean Tatar – L.S.Selendili) [5]. Recently, the problem of the preservation of Crimean Tatar has become even more pressing so the issues of the “pre-corpus” research of Crimean Tatar and the creation of its overt corpus are extremely urgent in terms of linguistics.

The main functions of the language are communicative and cognitive, however it is essential that you know how to handle “the construction material” of communication to be able to teach a human or a machine to communicate using the natural language [6]. The conducted research was based on the constructs – the syntactic units of different kinds which convey syntactic relations at the sentence level.

The key database of the research is the card-index which was composed by means of continuous sampling and computer cartography primarily on the material of scanned fiction texts (65 positions) and bilingual dictionaries [7]. A relatively low (for applied linguistics) number of sources are well explained by the restricted spheres of functioning of the Crimean Tatar language. The choice in favour of fiction texts for studying syntactic constructs is connected with the problem of norm breaking in Crimean Tatar and bilingual reflections which occur in the texts of other styles.

In the given paper the following terms are used:

1. The sentence is a syntactic construction, the environment in which the universal interaction of phenomena takes place; it combines sets of constructs and the linkup amongst them which causes the appearance of new properties and regularities that the constructs would not have if taken separately;

2. The syntactic constructions of Crimean Tatar are built by certain models out of syntactic constructs; the choice of syntactic constructs depends on the degree of simplicity/complexity and the communicative aims of the syntactic construction;

3. The basis of the syntactic constructs theory of the Crimean Tatar sentence is represented by the following notions:

– *a construct* – a speech unit which is manifested on the syntactic level and is used either by man or machine to create, realize, interpret, explain or restore the speech experience in terms of similarity and contrast;

– creating a real program of speech behaviour, a *construct* allows the communicator to explain someone’s verbal activity and to project their own verbal behaviour;

– *the verbal activity* of the recipient represents the organized process of implementation of multi-functional and multi-structural constructs by *the constructor*;

– *a constructor* is an actualizer of the complex of procedures, allowing the efficient manipulation of several objects at a time in the conditions of natural and artificial communication. To these belong: words, word combinations, speech patterns, addresses, coherent speech formulas, micro- and macrotext, etc.;

– *tools of a constructor* are a set of constructs, intended to model the sentences which consist of full sense words and normally of structural elements.

4. *The constructs* of the Crimean Tatar sentence are divided into the following groups: basic (general or stable), which comprise a rather wide scope of phenomena (sentences and their predicative parts, micro- and macro-texts, paragraphs, chapters, etc.) and auxilliary (stable or free), which have a rather narrow range of opportunities (regular words – autonomous and structural, all their possible combinations). Basic constructs convey the main information and auxilliary constructs may be changeable, having little effect on the main structure, but significant influence on the motivational context. The communicational meaning of stable constructs can be predicted while the communicational purpose and functions of free constructs depend on the constituents and may differ when the constituents are changed. It is important that the positions and meanings of auxilliary constructs be taken into account in the formal models of the Crimean Tatar sentence;

5. The identification of syntactic constructs in the Crimean Tatar sentence is the main condition for creating a specific communicative model of the language, bearing in mind its lexical-semantic and morphological peculiarities. To the syntactic constructs of Crimean Tatar belong: regular words (autonomous and structural), word combinations (subordinate and coordinative, predicative and non-predicative, stable and free, phraseologically and non-phraseologically inseparable), grammatical

cally bound and unbound syntactic formations (speech patterns, addresses and formulas of bound speech) and functional parts of integral constructions. Among them the following constructs are found: structural (regular words and stable word non-phraseological combinations in the function of subject or predicate), positional (words and word combinations in the functions of secondary and tertiary sentence members) and optional constructs of the sentence (speech patterns, formulas of bound speech and addresses);

6. The model of Crimean Tatar syntactic construction, which has symbols of the lexical semantic and morphological layout, is a formula which allows the communicator to reproduce and transform the sentence. It also allows the creation of a series of other syntactic constructions, substituting the elements of the same semantic or morphological group and filling out or compensating the gaps in the language models which are explained by the national specificity of the language, culture and the world outlook of its speakers;

7. Designing models of Turkic and Crimean Tatar sentences in huge massifs should be fulfilled bearing in mind the requirements concerning syntactic constructs, specificity and the algorithms of the automatic/semiautomatic syntactic layout of the text;

8. The creation of the theoretical basis for the engineering of appropriate programs of machine translation into/from poorly-known languages (such as Crimean Tatar) in the register of working languages, computer programs for studying, educational electronic programs and other products of modern informational culture-oriented technologies should rest on the pre-corpus processing of language datum;

9. Pre-corpus concepts on the system of constructs of the Crimean Tatar sentence, their nature and essence in the systematic linkups and relations of different layers, the peculiarities of their functioning and transformational options let us model the necessary syntactic construction thus practising the theory of syntactic constructs in the artificial environment. In scientific research the functions of such an artificial environment are carried out by the module “*Syntax and Phraseology*” of the computer lexicographic system “*The Russian Crimean Tatar Dictionary of Linguistic Equivalents*”.

The grammar of the Crimean Tatar language by A.Memetov has been the main source of pre-corpus processing [8]. In our work we took into account the research of G.Abdourakhmanov, A.Ablakov, Sh.S.Aylyarov, M.A.Askarova,

Y.D.Apresyan, I.Kh.Akhmatova, A.A.Baguirov, M.B.Balakayeva, A.N.Baskakov, I.P.Beletskaya, (Sevbo), V.V.Vinogradova, I.R.Vikhovanets, A.V.Dybo, M.Z.Zakiyev, V.P.Zakharov, G.A.Zolotova, V.I.Kormoushina, M.P.Kochergan, M.A.Kronhaus, A.Memetov, E.V.Rakhilina, N.K.Ryabtseva, V.A.Ploungian, Z.A.Sirazetdinov, E.I.Ubryatova, N.Chomsky and some others [9].

Modern computer technologies provide the opportunity to solve linguistic diffusive problems which are formulated empirically and do not have a complete solution by means of the mathematical explication of a linguistic object or phenomenon. On the one hand, such an approach allows the description of huge massifs of language material, and on the other hand, it creates conditions for the probabilistic prediction of language system functioning and demonstrates the combinatorial capabilities of those constructs which have not been described by the traditional grammar. In order to find out the structural models of Crimean Tatar constructs, we plan to devise the main parameters of their morphosyntactic and lexical semantic layout.

The syntactic layout establishes syntactic connections, the attribution of morphological and semantic characteristics to syntactic units, the discovery of the structural types of constructs which function in the sentence, and the determination of the syntactic function of constituents and the syntactic function of the constructs as members of the sentence.

Crimean Tatar shows some complexity in syntactic layout since there has been no evaluating system of constructs which can be used by a recipient to classify the different objects of their verbal space; the recipient may not always have the precise notion of the constructs that can be used to predict the frequently recurring events.

Making an attempt to describe the Crimean Tatar language in formal and applied aspects, it is necessary that attention be paid to the existing achievements in the spheres of structural, mathematical and applied linguistics and to the analysis of the main approaches in these spheres. Special attention should be paid to the linguistic support of the informational systems, e.g. to finding out which models, means and technologies are used to index the informational system on the basis of Russian, Turkic, Shor, Khakas, Bashkir, Tuvin, Kazakh and other languages [10].

In Russia, studies in the new field of corpus linguistics are developing productively with the support of the General Committee of the Russian Science Academy. A separate branch of this is the

language corpora of Russian peoples. Scientific research on the creation of corpora of minor Turkic languages is conducted under the guidance of A.V.Dybo and N.N.Shirobokova; the corpus of the ancient Turkic language is studied under the guidance of I.V.Kormoushina and I.A.Nevskaya [11].

“The Electronic Corpus of the Khakas language” has been created and allocated online [12]. The researching scholars have become the first in turcology to describe the principles of the automatic morphological analyzer for the Turkic languages (as an illustration, the version of the ancient Turkic language morphological analyzer is provided) and to extract its main components: the grammatical dictionary of the language; the ordinal model of the word form (the set of positions in the word form and morphonological notions of affixes in such positions); the rules of compatibility of affixes within the word form and two-level phonetic rules of the choice of allomorphs for a certain affix [13: 20-26]. The basic mechanism of the parser is the algorithm of analysis designed by F.Krylov on the basis of the STARLING system [14: 649-668].

Working on “The Electronic Corpus of the Ancient Turkic Texts”, a group of authors (I.V.Kormoushin, I.A.Nevskaya, A.V.Dybo, D.M.Nasylov, N.N.Shirobokova, and others) created a modern computer environment for the primary database on the basis of the STARLING multifunctional program complex. The project participants filled up the database and, using the STARLING program as well as *The Etymological Dictionary of Pre-Thirteenth-Century Turkish by G.Clauson* (Oxford, 1972), managed to compile an electronic ancient Turkic dictionary with a morphological and word-formation layout.

For the first time in history the foundations of the ancient Turkic corpus have been laid: a modern computer environment for the primary database has been created and the following texts have been digitalized and partly footnoted: “Maitrisimit”, the Uigur version of Syuan-Tzan’s biography (around 10000 word uses), the Brahmi writing texts (2800 word uses) and the Runic inscriptions of Altai-Sayan (3000 word uses). Expeditional investigations have provided the opportunity to clarify the set of signs of the Altaic Runic inscriptions and to find new ones, thus enhancing their database [15].

The corpus of the Shor language was created cooperatively with German scholars [16] with the fourth version of the Shoebox program which was invented to keep records and to help develop the literary form of different world non-literate languages and languages with recently acquired written forms.

The keyboard arrangement for the Bashkir language was designed in cooperation with Linux Inc. (Saint Petersburg) in the laboratory of linguistics and informational technologies, at the Language and Literature Institute of the Republic of Bashkortostan. The linguistic informational system “Machine Foundation of the Bashkir Language” was also created in the laboratory [17].

The work on the creation of “The Electronic Corpus of the Tuvin Language” started in 2011 in the Science Education Centre “Turkology” at Tuvin State University [18]. The designers of the corpus have transferred into electronic form and edited some of the prosaic works of the Tuvin writers of the Soviet and modern period, folklore texts, Tuvin verse, plays and official documents written in Tuvin and Russian (the Constitution of the Tuva Republic and several legislative documents on Parliament elections, the appointments of the officials, etc.) as well as the samples of folklore and everyday language of the Mongolian Tuvins. Also created, as part of the work, is the program “The search for morphemes in a given text”, which is written in the javascript language with the aim of looking for morphemes in Tuvin texts.

As for applied linguistics in Kazakhstan, this appeared with the research work of K.B.Bektayev. The data on the statistical informational typology of the Turkic text and the methods of algorithmic text processing, which were designed by K.B.Bektayev, give the opportunity to predict text structure and functions of the constructs, to conduct linguistic inspections and to uncover the calques and regularities in the production of the text [19]. Currently, supported by the Committee of Language Development and Social Political Work of the Kazakhstan Sport and Culture Ministry, the linguists and programmers have accomplished the major mission and “The Portal of the State Language” has been created [20].

The first steps in the sphere of Crimean Tatar structural, applied and mathematical linguistics have also been taken. Thus, R.Garabik, specialist of the Slovak National Corpus Department at the L.Shtur Institute of Linguistics of the Slovak Science Academy, and L.Kubedinova created “The Linguistic corpus of the Crimean Tatar Language” on the basis of publicistic texts [21].

While compiling the electronic “Dictionary of Russian Crimean Tatar Linguistic Correspondences” we juxtaposed the grammatical categories of the two different language structures necessary for creating the editor’s interface and administrator’s page of the dictionary. We determined the

main peculiarities of the grammatical categories present in both of the languages, the unique categories which exist only in one of the languages, gave the inventory data in the form of tables, fulfilled the planned cataloguing of the grammatical categories of the Russian and Crimean Tatar languages in respect of the synchronic aspect, classified the lexeme material, allocated the word groups with identical phonetic properties and built up the experimental samples which contain the grammatical parameters of words [22].

Pre-corpus grammar of the Crimean Tatar language represents data selection, organized according to certain rules, which is extracted out of the language system to exhibit the specific peculiarities of the language considered. Syntactic facts alongside the Crimean Tatar verbal data demonstrate the functioning and actualization of the constructs in the Crimean Tatar sentence within the language system. The syntactic constructions of the Crimean Tatar language are built up according to certain models out of syntactic constructs, their choice being dependent on the simplicity/complexity and communicative aim of syntactic construction.

In the capacity of the syntactic constructs of Crimean Tatar the following means are considered: autonomous and structural words, subordinate and coordinative word combinations, predicative and non-predicative word combinations, stable and free word combinations, phraseologically and non-phraseologically inseparable word combinations, speech patterns, formulas of bound speech, and addresses.

We appeal to the theory of syntactic constructs due to the fact that the traditional theory on the parts of the sentence leaves the gaps (lacunas) in the models of the sentence and the result of this is that some of the elements (addresses and speech patterns) that make up the sentence without any syntactic function are ignored. It is impossible to understand the full model of the Crimean Tatar sentence without accurate cataloguing. This means that we cannot use a separate model as a template for machine translation as a cliché to teach the Crimean Tatar language.

The pre-corpus processing of the correct, classical or literary Crimean Tatar texts allows us to describe the syntactic structure in detail, to extract immediate constituents in it and to determine the models of mathematical explication. It also enables us to formulate the notions of language-system and language-mechanism, the functioning of which is reflected in the verbal activity of its speakers, and

also to design the programs of machine translation and automatic text processing in the future. That is why we have made an attempt to draw a line between the Crimean Tatar compounds and word combinations, to discover the main models of compounds, to describe the word and its construction and semantic functions and to examine the lexical semantic peculiarities of structural words. We also aimed to uncover the role of the constructs in the formation of the motivational context and to demonstrate the semantic potential of the words inside the sentences, showing the associative relations in the words and their capabilities inside different thematic groups.

The order of words is the means of connection among structural constructs in the sentence, it binds the sentence with the motivational context, conveys the expressive emotional characteristics of the phrase, carries out the function of expressing the communicative grammatical meaning of the elements of the syntactic construction, and constitutes the manner of the materialization of lexical grammatical relations among the syntaxemes in the communicative composition.

The words, uniting with each other in strict sequences, form word combinations and the order of constructs constitutes specific constructions which render the finalized thought. Each part of the sentence has its unique function, morphosyntactic properties and communicative specificity. The opportunity of the communicator to choose among the primary (predicate and subject), secondary (attribute, object and adverbial modifiers which extend the primary parts of the sentence) and tertiary (attribute, object and adverbial modifiers which extend the secondary parts of the sentence) allows us to produce broad and narrow informational pieces which form a relatively finalized thought – the sentence [23: 184-185].

The data, which we have obtained on the functional peculiarities of the constructs of the Crimean Tatar sentence, are used in the creation of the module “Syntax and Phraseology” for the prototype of the lexicographical system “The Russian Crimean Tatar Dictionary of Linguistic Correspondences”. In illustration 1 there is the interface of the instrumental system of the electronic “Russian Crimean Tatar Dictionary of Linguistic Correspondences” – the page “Registry editor”, which is used to enlarge the main language registries. The Registry editor may work in two regimes: for the Russian and Crimean Tatar languages. The functionality of the given section allows us to enter new digits of the language registry and to give

them descriptions, naming the meanings of the corresponding grammatical, syntactic, and semantic parameters. The parameter “Part of speech spec.” apart from pointing to the part of speech, the definition of the semantic function of the construct (the main digit of the language entry). This parameter is used as the title of the construct for compiling the models of the sentences in the module “Syntax and Phraseology”.

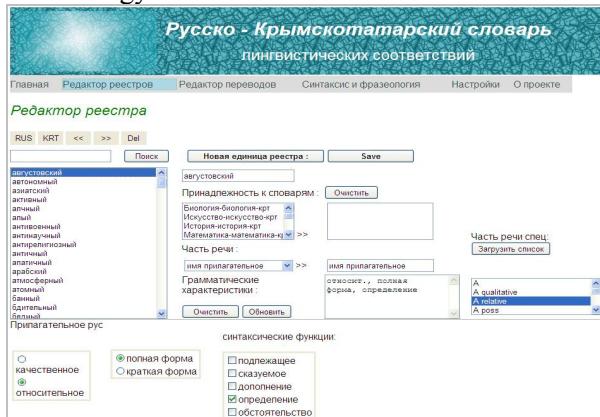


Illustration 1. Registry editor.

In order to represent the set of linguistic correspondences of the syntactic constructs and phraseological units in the electronic dictionary, the special editor “Syntax and Phraseology” was designed in the instrumental system of the electronic dictionary (Ill. 2).

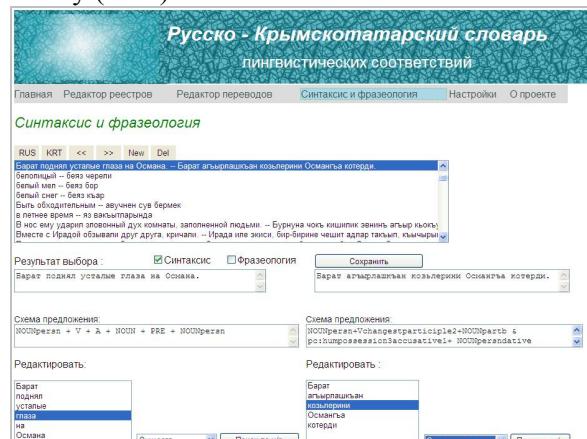


Illustration 2. Syntax and phraseology. Data processing in the “Syntax and Phraseology” module.

The access language to the Registry of the constructs and phraseological units is not only Russian but also Crimean Tatar. Each registry digit is flagged by the membership marker – “syntax” \diamond “phraseology” – and is later processed by different program modules, logged on to the given interface.

Data processing in the module “Syntax” is performed in the following way:

– a pair of syntactic constructs is chosen out of the registry or a new pair is introduced + the flag “Phraseology” is filled in + the “Save” operation is performed;

– if the layout of the sentences is already arranged, it will be displayed in the corresponding text fields where they can be edited.

Lower groups of the control elements for each language are used to compose the sentence layout. The sentence is divided into words and the search for a corresponding word form is accomplished in the main language registries of the dictionary and word forms. The search is available with or without referring to the part of speech. However, specifying the part of speech is supposed to decrease the time of the search. If the word form is found, its parameters are included in the sentence layout according to the template: 1) part of speech abbreviation + semantic function; 2) grammatical parameters of the word form. If the search has no results, certain information may be manually introduced into the sentence layout. Apart from composing the sentence layout, the “Syntax and Phraseology” module performs the indexing of the syntactic construct by the registry indexes of the concepts of which it is made up. This causes the situation when the dictionary entries of the concepts cumulate the syntactic examples of the performance of the given concept. These examples include the syntactic constructs of two languages (linguistic correspondence) and the models, which can be successfully implemented in teaching the languages.

The submodule “Phraseology” also performs indexing to demonstrate the phraseological correspondences in the dictionary entries of the concepts of the main language registries. The instrumental system of the electronic dictionary is created as the web application (C#, Java), integrated with the SQL database Server. At this stage the instrumental system and the lexicographic database of the dictionary are working in the test regime in the Science Research Centre of the Crimean Tatar language, literature and history named after Bekir Choban-zadeh.

The present article, for the first time, describes the categorical properties of the grammatically-bound and not-bound constructs of the Crimean Tatar sentence and the main factors of their functioning and formation, and considers the processes of the interaction between the lexical and grammatical semantics of the components of every construct taken apart and inside the sentence. Among the results of the work described in the article the following may be mentioned: models have been

developed of the syntactic constructs, taking into account the parameters of the morphosyntactic and lexical semantic layout; the transformational capabilities and the communicative, lexical semantic, psycholinguistic and pragmatic peculiarities of the Crimean Tatar sentence have been revealed, consisting of structural and informational, logical and factual constructs; and the fundamental theory of the Crimean Tatar word combination has been proposed. In addition, and for the first time, the module “Syntax and Phraseology” has been developed for “The Russian Crimean Tatar Dictionary of Linguistic Correspondences”, taking into account the parameters of the morphosyntactic and lexical semantic layout. This allows us to create models of the syntactic constructs on the basis of parallel texts in the semi-automatic regime.

References

1. Akmollaev E.S. Classificatsiya bessoyuznykh slozhnykh predlozhenii i osobennosti ikh sootnosheniya s tipami soyuznykh predlozhenii: (na materiale bessoyuznykh slozhnykh predlozhenii otkrytoi struktury): avtoref.dis...kand.philol.nauk: 10.02.08. Tashkent, TGPI, 1986. (in Russian)
2. Rakhilina E.V. Lingvistika konstruktseii / Pod red. Otvetstvennyi redaktor E.V.Rakhilinoi, T.I.Reznikovoi. Moskva: Azbukovnik, 2010. 584 p. (in Russian)
3. Emirova A.M. Krymskotatarskaya leksikografiya: sovremennoe sostoyanie i perspektivy razvitiya [Elektronnyi resurs] // Kul'tura narodov Prichernomor'ya: Nauchnyi zhurnal. Simferopol', 1997. №3-sentjabr'. Rezhim dostupa: <http://turkology.tk/library/161> (date of use 10.11.2014). (in Russian)
4. Shcherba L.V. Yazykovaya sistema i rechevaya deyatel'nost' / L.V.Shcherba. Izd. 2-e, stereotipnoe. M.: Editorial URSS, 2004. P. 265-304; Guerd A.S., Ivashko L.A., Lutovinova I.S. i dr. Osnovnye tipy slovarei v otechestvennoi rusistike // Leksikografiya russkogo yazyka. Uchebnik dlya vysshikh uchebnykh zavedenii. SPb.: Fakultet filologii i iskusstv SPbGU, 2009. 672 p. (in Russian)
5. Yazyki mira, nahodyashchiesya pod ugrozoy ischeznoveniya [Elektronnyi resurs]. Rezhim dostupa: <http://www.unesco.org/new/ru/culture/themes/endangered-languages/atlas-of-languages-in-danger/> (date of use 10.11.2014). (in Russian)
6. Vinogradov V.V. Iz istorii izucheniya russkogo sintaksisa. M.: Izd-vo Mosk. Un-ta, 1958. 400 p.; Grammatika russkogo yazyka / redkol.: akad. V.V.Vinogradov, chl.-kor. AN SSSR E.S.Istrina. M.: Izd-vo Akad. nauk SSSR, 1954. T. 2 : Sintaksis. Ch. 1. 703 p. (in Russian)
7. Abdullaev E.M. Russko-krymskotatarskii uchebnyi slovar': bolee 5000 slov / E.M.Abdullaev, M.U.Umerov. Simferopol' : Krymchpedgiz, 1994. 384 p.; Kratkiy slovar' kognitivnykh terminov / E.S.Kubryakova [i dr.]. M.: Fil. fak. MGU, 1997. 245s.; Konstrukt // Filosofskii slovar'. Biblioteka "Polka bukinista". Znachimye knigi otechestvennykh i zarubezhnykh avtorov [Elektronnyj resurs] Rezhim dostupa: <http://philosophy.polbu.ru/konstrukt.htm> (date of use 10.11.2014). (in Russian)
8. Memetov A.M. Zemanevii k'rymtatar tili. Simferopol: K'rym devlet ok'uv pedagogika neshriyati, 2006. 320 p. (K'rymtatar tilinde); Memetov A.M. Krymtatarskii yazik. Part. 1. Obshchiye svedeniya o yazike; Part. 2. Morphologiya: ucheb. posobiye / A.M.Memetov, K.M.Musayev. Simferopol: Krymchpedgiz, 2003. 287, [1] p. (in Crimean Tatar)
9. Buskunbaeva L.A., Siraztdinov Z.A. K sisteme razmetok v natsional'nom korpusse bashkirskogo yazyka // Aktual'nye problemi dialektologii yazykov narodov Rossii. Materialy XI mezhregional'noy konferencii. Ufa. S. 50-55; Dybo A.V., Sheymovich A.V. Avtomaticheskii morfologicheskii analiz dlja korpusov tyurkskikh yazykov // Filologiya i kul'tura, 2014. №2(36). S.20-26; Zakharov V.P. Korpusnaya lingvistika: uchebno-metodicheskoe posobie / V.P.Zakharov. Sankt-Peterburg: Sankt-Peterburgskii gos. universitet, 2005. 48 p.; Korpus po pamyatnikam runicheskogo pis'ma Gornogo Altaja [Elektronnyj resurs]. Rezhim dostupa: <http://www.altay.uni-frankfurt.de/> (date of use 10.11.2014); Korpusy po pamyatnikam tyurkskikh yazykov [Elektronnyj resurs]. Rezhim dostupa: http://www.tuvancorpus.ru/?q=korpusy_po_pamyatnikam_tyurkskikh_yazykov (date of use 10.11.2014); Ploungyan V.A. Nacional'nyi korpus russkogo yazyka: obshaya harakteristika / Ploungyan V.A., Reznikova T.I., Sichinava D.V. // NTI. Seriya 2. 2005. № 3. P. 9-13; Ploungyan V.A. Nacional'nyi korpus russkogo yazyka kak instrument leksikografa / V.A.Ploungyan, D.V.Sichinava // Slovo i slovar' = Vocabulum et vocabularium: cb. nauch. tr. po leksikografii [Rychkova L.V., Voronovich V.L., Emelyanova S.A. (otv. red.)]. Grodno: GrGU, 2005. P. 197-202; Ploungyan V.A. Nacional'nyi korpus russkogo yazyka: opyt sozdaniya korpusov tekstov sovremennogo russkogo yazyka / V.A.Ploungyan, D.V.Sichinava // Trudy Mezhd. konferencii "Korpusnaya lingvistika-2004" [L.N.Belyaeva i dr. (red.)]. SPb: SPbGU, 2004. P. 216-238. (in Russian, Ukrainian)
10. Bektaev K.B. Statistiko-informacionnaya tipologiya tyurks'kogo teksta : avtoref. dis. na zdobuttya nauk. stupenya d. filol. nauk: spec. 10.02.21: Strukturnaya, prikladnaya i matematicheskaya lingvistika / Bektaev Kaldybay Bektaevich; Akademiya nauk SSSR, Institut yazykoznanija, Leningradskoe otdelenie. – Leningrad, 1975. 39 p.; Dybo A.V., Sheymovich A.V. Avtomaticheskii

- morfologicheskii analiz dlya korpusov tyurkskikh yazykov // Filologiya i kul'tura, 2014. №2(36). P.20-26; *Zakharov V.P.* Korpusnaya lingvistika : uchebno-metodicheskoe posobie / V.P.Zakharov. Sankt-Peterburg: Sankt-Peterburgskii gos. universitet, 2005. 48 s.; Nacional'nyi korpus russkogo yazyka [Elektronnyi resurs]. Rezhim dostupa: <http://www.ruscorpora.ru> (date of use 10.11.2014); Korpus po pamyatnikam runicheskogo pis'ma Gornogo Altaja [Elektronnyi resurs]. Rezhim dostupa: <http://www.altay.uni-frankfurt.de/> (date of use 10.11.2014); Korpusna lingvistika: monografiya / V.A.Shirov [ta in.]; NAN Ukrainsk, Ukr. mov.-inform. fond. Kiiv: Dovira, 2005. 472 p.; Korpusy po pamyatnikam tyurkskikh yazykov [Elektronnyi resurs]. Rezhim dostupa: http://www.tuvancorpus.ru/?q=korpusy_po_pamyatnikam (date of use 10.11.2014) tyurkskikh_yazykov i mn. dr. (in Russian, Ukrainian)
11. *Dybo A.V., Sheymovich A.V.* Avtomaticheskii morfologicheskii analiz dlya korpusov tyurkskikh yazykov // Filologiya i kul'tura, 2014. №2(36). P.20-26; *Sheymovich A.V.* Morfologicheskaya razmetka korpusa hakasskogo yazyka // Rossiyskaja tyrkologiya. № 2(5). S. 48-61. (in Russian)
12. Elektronnyi korpus hakasskogo yazyka [Elektronnyi resurs]. Rezhim dostupa: <http://khakas.altaica.ru/grammar/> (date of use 10.11.2014). (in Russian)
13. *Dybo A.V., Sheymovich A.V.* Avtomaticheskii morfologicheskii analiz dlya korpusov tyurkskikh yazykov // Filologiya i kul'tura, 2014. №2(36). P.20-26.
14. *Krylov S.A.* Strategii primeneniya integrirovannoj informacionnoj sredy StarLing v korpusnoi lingvistike i v kompjuternoi leksikografii // Orientalia et classica. Trudy Instituta vostochnykh kul'tur i antichnosti. Vypusk XIX. Aspekyt komparativistiki. 3. M., RGGU, 2008. P. 649-668. (in Russian)
15. Korpus po pamyatnikam runicheskogo pis'ma Gornogo Altaja [Elektronnyi resurs]. Rezhim dostupa: <http://www.altay.uni-frankfurt.de/> (date of use 10.11.2014); Korpusy po pamyatnikam tyurkskikh yazykov [Elektronnyj resurs]. Rezhim dostupa: http://www.tuvancorpus.ru/?q=korpusy_po_pamyatnikam_tyurkskikh_yazykov(date of use 10.11.2014); Elektronnyi korpus po pamyatnikam doislamskikh drevnetyurkskih tekstov [Elektronnyi resurs]. Rezhim dostupa: <http://vatec2.fkidg1.uni-frankfurt.de/> (date of use 10.11.2014). (in Russian)
16. Elektronnyi korpus shorskogo yazyka [Elektronnyi resurs]. Rezhim dostupa: <http://shoriya.ngpi.rdtc.ru> (date of use 10.11.2014). (in Russian)
17. Mashinnyy fond bashkirskogo yazyka [Elektronnyi resurs]. Rezhim dostupa: www.mfbl.ru (date of use 10.11.2014). (in Russian)
18. *Salchak A.Y.* Elektronnyi korpus tuvinskogo yazyka [Elektronnyi resurs]/ Mezhdunarodnaya nauchno-prakticheskaya konferenciya, posvyashchennaya 100-letiyu so dnja rozhdeniya "Narodnogo akademika" Vladimira Mihailovicha Nadelyaeva. Kyzyl // Novye issledovaniya Tuvy. Rezhim dostupa: http://www.tuva.asia/journal/issue_15/5231-salchak.html (date of use 10.11.2014). (in Russian)
19. *Bektaev K.B.* Statistiko-informacionnaya tipologiya tyurks'kogo teksta : avtoref. dis. na zdobuttya nauk. stupenya d. filol. nauk: spec. 10.02.21: Strukturnaya, prikladnaya i matematicheskaya lingvistika / Bektaev Kaldybay Bektaevich; Akademija nauk SSSR, Institut yazykoznanija, Leningradskoe otdelenie. – Leningrad, 1975. 39 p. (in Russian)
20. Portal gosudarstvennogo yazyka [Elektronnyi resurs]. Rezhim dostupa: <http://til.gov.kz/wps/portal/> (date of use 10.11.2014). (in Russian)
21. Lingvisticheskij korpus krymskotatarskogo yazyka [Elektronnyy resurs]. Rezhim dostupa: <http://korpus.juls.savba.sk/QIRIM/#id9> (date of use 10.11.2014). (in Russian)
22. *Okaz L.S.* Sopostavitel'naya tipologiya prichastii krymskotatarskogo i russkogo yazykov / L.S.Okaz, V.N.Alieva // Uchenye zapiski Tavricheskogo nacional'nogo universiteta im. V.I.Vernadskogo. Seriya "Filologiya. Social'nye kommunikacii". 2009. T. 22(61), №3. P.46-50; *Okaz L.S.* Tipologicheski znachimiye aspekyt sopostavleniya grammaticeskikh kategorii krymskotatarskogo i russkogo yazykov (imennye chasti rechi) / L.S.Okaz, V.N.Alieva // Kul'tura narodov Prichernomor'ya. 2008. №147. T.2. S.96-99. (in Russian)
23. *Okaz L.S.* Sredstva vyrazheniya sintaksicheskikh otnoshenii mezhdu chastyami slozhnogo predlozheniya v krymskotatarskom yazyke / L.S.Okaz // Kul'tura narodov Prichernomor'ya. 2001. №17. S.184-185. (in Russian)

ДОКОРПУСНАЯ ОБРАБОТКА КРЫМСКОТАРСКИХ ТЕКСТОВ

Лемара Сергеевна Селендили,

Московский государственный гуманитарный университет имени М.А. Шолохова,
Россия, 109240, г.Москва, ул.Верхняя Радищевская, д.16-18,
lemara2002@hotmail.com.

Статья посвящена описанию теории синтаксических конструктов на материале крымскотатарского языка. Докорпусная обработка языкового материала осуществляется с целью создания теоретической базы для разработки адекватных систем машинного перевода с/на крымскотатарский язык академических электронных словарей и инструментальных систем, направленных на развитие языка и практики преподавания языков с помощью современных информационных технологий. В процессе конструирования моделей на основе больших объемов данных учитываются условия и последовательность автоматической / полуавтоматической синтаксической разметки текста. Средой реализации синтаксических конструктов является модуль «Синтаксис и фразеология», созданный для базы лексикографических данных «Русско-крымскотатарского словаря лингвистических соответствий».

Ключевые слова: крымскотатарский язык, синтаксис, структурная, прикладная и математическая лингвистика, лингвистический корпус, докорпусная обработка текста.

Синтаксис как одна из фундаментальных лингвистических наук является неотъемлемой частью теории любого языка. Это очень важный участок для языков, которые относительно недавно стали развивать свои теоретические основы. Среди них, безусловно, следует отметить крымскотатарский язык, научному изучению синтаксиса которого (кроме наших работ) посвящено, насколько нам известно, только одно диссертационное исследование – кандидатская диссертация Э.С.Акмоллаева [1].

Создание синтаксической теории конструктов крымскотатарского предложения приобретает особое значение в свете современной универсальной теории «Лингвистики конструкций» Е.В.Рахилиной, которая «ориентирована на изучение и описание реального естественного языка, причем в любых аспектах и срезах и на любых уровнях» и предназначена «для практики: словарной, грамматической, практики корпусного анализа, экспедиционных работ и типологических исследований» [2: 74], имеет значение для прикладной лингвистики, основанной на формализации информации о языковых объектах: план выражения и план содержания синтаксических единиц, их морфологические, семантические и синтаксические особенности, формулы, определенные совокупности знаков (символов). Такой подход требует точности описания, «поскольку конструкция – сложный лингвистический объект, в ее перестройке могут участвовать все ее компоненты, даже если они являются единицами разных уровней» [2: 538]. «Для этого, бесспорно, луч-

ше опираться на конструкт, а не на естественное словоупотребление, которое, конечно, далеко от формального» [2: 238].

Лингвистическая теория активно применяется в создании компьютерных учебных программ на разных языках, являющихся для пользователей как родными, так и иностранными, разработке различных лексикографических систем, лингвистических корпусов, средств автоматической обработки естественного языка, обеспечивающих обмен информацией, общение человека с машиной на естественном языке.

Особенно активно в последнее время развиваются технологии корпусной лингвистики. Ученые поднимают проблемы автоматической семантической и синтаксической разметки текстов лингвистического корпуса и пытаются найти пути их решения. Кроме того, для создания программ машинного перевода все чаще проводятся корпусные лингвистические исследования на основе параллельных текстов, создаются многоязычные электронные словари и, что крайне важно, апробируются грамматики конкретных языков.

Важно отметить, что не все тюркские языки имеют академические грамматики, современные словари (не только электронные, но и, прежде всего, бумажного типа), поэтому некоторые из них нуждаются еще и в «докорпусной» обработке.

Так, лексикография и синтаксис крымскотатарского языка изучены весьма фрагментарно. А.М.Эмирова отмечает, что «существующие сегодня словари очень ярко иллюстрируют бед-

ственное состояние крымскотатарского языко-знания» и что «только на базе представительных словарей, отражающих все параметры языковой системы, возможно объективное научное описание и изучение какого-либо языка» [3]. Нет современных словарей академического типа, энциклопедических словарей, тезаурусов, идеографических словарей, толкового словаря и др. [4], отсутствуют детализированные данные о структуре синтаксической конструкции, не изучались особенности функционирования словосочетаний, до сих пор не определено место речевых формул, обращений, остается нерешенным вопрос о членах крымскотатарского предложения; «ждут своего исследования с применением современного концептуально-терминологического аппарата такие участки языковой системы, как фонетика, фонология, грамматика, лексикология, фразеология, лексикография и др.» [3].

Согласно данным Центра новостей ООН со ссылкой на обновленную в декабре 2010 года информацию интерактивного атласа ЮНЕСКО «Языки мира, находящиеся под угрозой исчезновения», крымскотатарский язык находится под угрозой исчезновения, особенно его ногайский диалект (северный степной диалект крымскотатарского языка – Л.С. Селендили) [5], в последние годы проблема сохранения крымскотатарского языка обострилась еще больше, поэтому проблемы «докорпусного» исследования крымскотатарского языка и создания открытого корпуса крымскотатарского языка стоят перед лингвистической наукой как сверхактуальные.

Важнейшими функциями языка являются когнитивная и коммуникативная, но, чтобы научить субъекта (человека или машину) общаться на естественном языке, необходимо знать, как обращаться со «строительным материалом» этого общения [6]. В проведенном научном исследовании в качестве «строительного материала» коммуникации определены конструкты – синтаксические единицы различного характера, которые являются носителями синтаксических отношений на уровне предложения.

Ключевой базой исследования является картотека, составленная методом сплошной выборки и компьютерной картографии преимущественно из сканированных художественных текстов (65 позиций) и двуязычных словарей [7]. Относительно небольшой для прикладной лингвистики объем источников базы объясняется ограниченностью сфер функционирования крымскотатарского языка. Выбор

именно художественных текстов для исследования синтаксических конструктов связан с проблемой нарушения норм крымскотатарского языка и билингвальными проявлениями, которые имеют место в текстах других стилей.

В настоящей работе используются следующие термины:

1. Предложение – это синтаксическая конструкция, среда, в которой происходит всеобщее взаимодействие явлений, наблюдается вхождение части в целое, выражаются отношения между совокупностью конструктов и связью, объединяющей эти конструкты, что приводит к появлению у совокупности новых свойств и закономерностей, не присущих конструктам в их разобщенности.

2. Синтаксические конструкции крымскотатарского языка строятся по определенным моделям из синтаксических конструктов; выбор синтаксических конструктов зависит от простоты / сложности и коммуникативной предназначенности синтаксической конструкции.

3. Основу теории синтаксических конструктов крымскотатарского предложения представляют следующие понятия:

- *конструкт* – единица речи, реализуемая на синтаксическом уровне, которую человек или машина использует для того, чтобы создать, осознать или истолковать, объяснить или предсказать, воссоздать речевой опыт в терминах схожести и контраста;

- задавая фактическую программу речевого поведения, *конструкт* позволяет коммуниканту объяснить чужую речевую деятельность и проектировать собственное речевое поведение;

- *речевая деятельность* реципиента представляет собой организованный процесс использования *конструктором* системы разноструктурных и разнофункциональных конструктов;

- *конструктор* – это реализатор комплекса процедур, позволяющий умелое одновременное манипулирование несколькими объектами, такими как слова, словосочетание, речевые формулы, обращение, формулы связной речи, микро- и макротекст и т.д. в условиях естественной и искусственной коммуникации;

- *инструментом конструктора* является набор конструктов, предназначенный для моделирования предложений, состоящих из совокупности полнозначных слов и, как правило, структурных элементов.

4. *Конструкты* крымскотатарского предложения делятся на следующие группы: базо-

вые (общие или устойчивые), которые включают в себя относительно широкий спектр явлений (предложения и их предикативные части, микро- и макротексты, абзацы, главы и т.п.) и вспомогательные (устойчивые или свободные), имеющие узкий диапазон возможностей (слова – знаменательные и структурные, их всевозможные комбинации и сочетания). Базовые конструкты передают основную информацию, вспомогательные конструкты могут заменяться, незначительно изменяя основную структуру, но меняя мотивационный контекст. Коммуникативное значение устойчивых конструктов можно прогнозировать, а коммуникативное предназначение и функции свободных конструктов зависят от непосредственно составляющих и от их замены могут изменяться. В формальных моделях крымскотатарского предложения важно учитывать позиции и значение вспомогательных конструктов.

5. Идентификация синтаксических конструктов крымскотатарского предложения – это условие создания конкретной коммуникативной модели языка с учетом лексико-семантических и морфологических особенностей. К синтаксическим конструктам крымскотатарского языка относятся слова (полнозначные и структурные), словосочетания (подчинительные и сочинительные, предикативные и непредикативные, устойчивые и свободные, устойчивые и неделимые, устойчивые фразеологизированного и нефразеологизированного характера), грамматически связанные и грамматически несвязанные синтаксические образования (речевые формулы, обращение, формулы связной речи), функциональные части целостных конструкций. Среди них выделяются структурные (слова, устойчивые словосочетания нефразеологизированного характера в функции субъекта или предиката), позиционные (слова, словосочетания, сочетания слов в функции второстепенных и третьестепенных членов предложения) и факультативные конструкты предложения (речевые формулы, формулы связной речи, обращение).

6. Модель крымскотатарской синтаксической конструкции, содержащей символы лексико-семантической и морфологической разметки, является формулой, позволяющей коммуниканту воспроизводить, трансформировать предложение, создавать множество других синтаксических конструкций, подставляя элементы одной и той же семантической или морфологической группы, заполнять или компенсировать

лакуны (пробелы) в языковых моделях, обусловленные национальной спецификой языка, культуры и мировоззрения его носителей.

7. Проектирование моделей тюркских и крымскотатарских предложений на больших массивах должно осуществляться с учетом требований, предъявляемых к синтаксическим конструктам, специфики и алгоритмов автоматической / полуавтоматической синтаксической разметки текста.

8. Создание теоретической базы для проектирования адекватных программ машинного перевода на/с малоизученных языков (к которым относится крымскотатарский) в реестре рабочих языков, компьютерных учебных программ, учебных электронных словарей и других продуктов современных информационных культурно-ориентированных технологий должно базироваться на докорпусной обработке языкового материала.

9. Докорпусные представления о системе конструктов крымскотатарского предложения, их природе и сущности в разноуровневых системных связях и отношениях, об особенностях их функционирования и трансформационных возможностях позволяют смоделировать необходимую синтаксическую конструкцию и практически реализовать теорию синтаксических конструктов в искусственной среде. В научном исследовании такой искусственной средой стал модуль «Синтаксис и фразеология» компьютерной лексикографической системы «Русско-крымскотатарский словарь лингвистических соответствий».

Фундаментом докорпусного описания явилась грамматика крымскотатарского языка А.Меметова [8]. В работе мы опирались на исследования Г.Абдурахманова, А.Аблакова, Ш.С.Айлярова, М.А.Аскаровой, Ю.Д.Апресяна, И.Х.Ахматова, А.А.Багирова, М.Б.Балакаева, А.Н.Баскакова, И.П.Белецкой (Севбо), В.В.Виноградова, И.Р.Выхованца, А.В.Дыбо, М.З.Закиева, В.П.Захарова, Г.А.Золотовой, В.И.Кормушина, М.П.Кочергана, М.А.Кронгауза, А.Меметова, Е.В.Рахилиной, Н.К.Рябцевой, В.А.Плунгяна, З.А.Сиразитдинова, Е.И.Убрайтовой, Н.Хомского и многих других [9].

Современные компьютерные технологии позволяют разрешать диффузные, эмпирически сформулированные и не имеющие полного решения лингвистические задачи математической экспликацией лингвистического объекта или явления. С одной стороны, такой подход позволяет описывать большие массивы языкового

материала, с другой – он создает условия вероятностного прогнозирования функционирования языка как системы и демонстрирует комбинаторные возможности конструктов, которые не были описаны традиционной грамматикой. С целью выявления структурных моделей конструктов крымскотатарского языка мы планируем разработать параметры морфосинтаксической и лексико-семантической разметки.

Синтаксическая разметка представляет собой фиксацию синтаксических связей, приписывание синтаксическим единицам соответствующих морфологических и семантических характеристик, выявление структурных типов конструктов, функционирующих в предложении, определение синтаксической функции непосредственно составляющих, обозначение функции синтаксических конструктов как членов предложения.

Для крымскотатарского языка существуют определенные сложности синтаксической разметки, потому что до сих пор не было оценочной системы конструктов, которая используется реципиентом для классификации различных объектов его речевого пространства; реципиент не всегда имеет узкое представление о том, какие конструкты использовать для прогнозирования повторяющихся событий.

Предпринимая попытку описания крымскотатарского языка в формальном и прикладном аспектах, важно обратить внимание на существующие достижения в сфере структурной, математической и прикладной лингвистики, проанализировать подходы, развитые в этой области мировой наукой. Особое внимание следует обратить на лингвистическое обеспечение информационных систем: например, какими моделями, средствами и технологиями индексирована информационная система на базе русского, турецкого, шорского, хакасского, башкирского, тувинского, казахского и других языков [10].

В России при поддержке Президиума РАН продуктивно развиваются исследования по новой программе: корпусная лингвистика. Отдельное направление этой программы – корпуса языков народов России. Под руководством А.В.Дыбо и Н.Н.Широбоковой проводятся научные исследования по созданию корпусов моноритарных тюркских языков, И.В.Кормушина и И.А.Невской – корпуса древнетюркского языка [11].

Создан и размещен в сети Интернет «Электронный корпус хакасского языка» [12]. В про-

цессе работы над корпусом исследователями впервые в тюркологии описаны принципы работы автоматического морфологического анализатора для тюркских языков (в качестве иллюстрации приводится версия морфологического анализатора для древнетюркского языка), выделены его основные компоненты: грамматический словарь языка; порядковая модель словоформы (набор позиций в словоформе и морфонологических представлений аффиксов для этих позиций); правила сочетаемости аффиксов в пределах словоформы и двухуровневые фонетические правила выбора алломорфов конкретного аффикса [13: 20-26]. В основе работы парсера лежит алгоритм анализа, разработанный Ф.Крыловым на базе системы Starling [14: 649-668].

В процессе работы над созданием «Электронного корпуса древнетюркских текстов» коллективом авторов (И.В.Кормушин, И.А.Невская, А.В.Дыбо, Д.М.Насилов, Н.Н.Широбокова и др.) на базе многофункционального программного комплекса Starling создана современная компьютерная среда для первичной базы данных; участники проекта заполнили базу данных и на базе программы Starling, а также словаря «Clauson G. An Etymological Dictionary of Pre-Thirteenth-Century Turkish. Oxford, 1972» проделали работу по составлению электронного древнетюркского словаря с морфологической и словообразовательной разметкой.

Впервые в тюркологии заложены основы древнетюркского корпуса: создана современная компьютерная среда для первичной базы данных, дигитализированы и частично аннотированы тексты «Майтрасимит» и Уйгурская версия биографии Сюань-Цзана уйгурского письма (около 10000 словоупотреблений), тексты на письме брахми (2800 словоупотреблений), а также тексты рунических надписей Алтая-Саяна (3000 словоупотреблений). Экспедиционные исследования позволили уточнить состав знаков алтайских рунических надписей и открыть новые надписи, пополнив их базу данных [15].

С применением четвертой версии программы Shoebox, разработанной с целью документации и помощи в развитии литературной формы бесписьменных и младописьменных языков народов мира, совместно с немецкими учеными был создан корпус шорского языка [16].

В лаборатории лингвистики и информационных технологий, созданной при ИИЯЛ Рес-

публики Башкортостан, выполнены работы с компанией Линукс Инк (Санкт-Петербург) по созданию клавиатурной раскладки для башкирского языка, создана лингвистическая информационная система «Машинный фонд башкирского языка» [17].

В Научно-образовательном центре «Тюркология» Тувинского государственного университета с 2011 года началась работа по созданию «Электронного корпуса текстов тувинского языка» [18]. Разработчиками переведены в электронный вид и отредактированы некоторые прозаические произведения тувинских писателей советского периода, писателей современного периода, поэтические тексты, фольклорные тексты, пьесы, тексты официально-деловых документов на тувинском и русском языках (Конституция Республики Тыва и некоторые законодательные документы о выборах депутатов, должностных лиц и т.д.), а также образцы фольклора и бытовой речи тувинцев Монголии. Создана программа «Поиск морфем в заданном тексте» на языке программирования javascript, предназначенная для поиска морфем в текстах на тувинском языке.

Прикладная лингвистика в Казахстане началась с исследований К.Б.Бектаева. Сведения о статистико-информационной типологии тюркского текста и методы алгоритмизированной обработки текста, разработанные К.Б.Бектаевым, позволяют прогнозировать структуру текста, функции конструктов, выполнять лингвистическую экспертизу, выявлять кальки и закономерности производства текста [19]. В настоящее время при поддержке Комитета по развитию языков и общественно-политической работы Министерства культуры и спорта Республики Казахстан лингвистами и программистами проделана большая работа: создан «Портал государственного языка» [20].

Первые шаги предприняты и в области крымскотатарской структурной, прикладной и математической лингвистики. Так, Р.Гарабик, специалист отдела Словацкого национального корпуса Института языкоznания имени Л.Штура Словацкой Академии наук, и Л.Кубединова на материале публицистических текстов создали «Лингвистический корпус крымскотатарского языка» [21].

В процессе создания электронного «Русско-крымскотатарского словаря лингвистических соответствий» нами сопоставлены грамматические категории двух разноструктурных языков, необходимые для создания интерфейса редак-

тора, страницы администратора электронного словаря, определены грамматические категории, имеющиеся в обоих языках, и обозначены их особенности, выявлены грамматические категории, характерные только для одного из сопоставляемых языков, представлены инвентарные данные в виде таблиц, осуществлена планомерная инвентаризация грамматических категорий русского и крымскотатарского языков в синхронном аспекте, классифицирован лексемный материал, выделены группы слов с одинаковыми фонетическими признаками, созданы экспериментальные образцы, содержащие грамматические параметры слов [22].

Докорпусная грамматика крымскотатарского языка представляет собой сформированную по определенным правилам выборку данных из области реализации языковой системы, которая содержит специфические особенности рассматриваемого языка. Синтаксические факты с использованием крымскотатарского речевого материала демонстрируют функционирование и реализацию конструктов крымскотатарского предложения внутри языковой системы. Синтаксические конструкции крымскотатарского языка строятся по определенным моделям из синтаксических конструктов; выбор синтаксических конструктов зависит от простоты / сложности и коммуникативной предназначенностии синтаксической конструкции.

В качестве синтаксических конструктов крымскотатарского языка рассматриваются полнозначные и служебные слова, подчинительные и сочинительные словосочетания, предикативные и непредикативные словосочетания, свободные и устойчивые словосочетания, устойчивые словосочетания фразеологизированного и нефразеологизированного характера, речевые формулы и формулы связной речи, обращение.

Мы обращаемся к теории синтаксических конструктов в связи с тем, что традиционная теория о членах предложения оставляет лакуны в моделях предложения: некоторые из элементов (например, обращение, речевые формулы), представляющие структуру предложения, но не выполняющие конкретную синтаксическую функцию, остаются неучтенными. Без четкой инвентаризации всех компонентов представить полную модель крымскотатарского предложения невозможно. Значит, невозможно использовать отдельно взятую модель как шаблон для машинного перевода, как клише для обучения крымскотатарскому языку.

Докорпусная обработка правильных, классических или соответствующих литературной норме крымскотатарских текстов позволяет детально описывать синтаксическую структуру, выделять в ней непосредственно составляющие, выявлять модели математической экспликации, формулировать представления о языке-системе и языке-механизме, функционирование которого проявляется в речевой деятельности его носителей, для создания в будущем программ машинного перевода и автоматической обработки текста. Поэтому мы предприняли попытку разграничить крымскотатарские сложные слова и словосочетания, выявить модели построения сложных слов, описать слово и его конструктивно-семантические функции, рассмотреть лексико-семантические особенности структурных слов, выявить роль слов-конструктов в формировании мотивационного контекста, продемонстрировать семантический потенциал слов внутри предложений, показать ассоциативные связи слов, их возможности в границах разных тематических групп.

Порядок слов в предложении является средством связи между структурными конструктами предложения, средством реализации связи предложения с мотивирующим контекстом, средством выражения эмоционально-экспрессивной характеристики высказывания, выполняет функцию выражения коммуникативно-грамматического значения элементов крымскотатарской синтаксической конструкции, является способом реализации лексико-грамматических отношений синтаксем в коммуникативной композиции.

Слова, объединяясь между собой в строгой последовательности, образуют словосочетания, порядок расположения конструктов формирует конструкцию, передающую законченную мысль. Каждый член предложения имеет свою функцию, синтаксические признаки и коммуникативную специфику. Возможность права выбора в использовании коммуникантом главных (сказуемое и подлежащее), второстепенных (определение, дополнение и обстоятельство, распространяющие главные члены предложения), третьестепенных членов предложения (определение, дополнение и обстоятельство, распространяющие второстепенные члены предложения) позволяет создавать широкие и узкие информационные отрезки, формирующие относительно законченную мысль – предложение [23: 184-185].

Полученные нами данные о функциональных особенностях конструктов крымскотатарского предложения использованы при создании модуля «Синтаксис и фразеология» для прототипа лексикографической системы «Русско-крымскотатарский словарь лингвистических соответствий». На рис.1 представлен интерфейс инstrumentальной системы электронного «Русско-Крымскотатарского словаря лингвистических соответствий» – страница «Редактор реестра», предназначенная для пополнения основных языковых реестров. Редактор реестра может работать в двух режимах: для крымскотатарского языка и русского. Функциональность данного раздела позволяет вводить новые единицы языкового реестра и давать им описание, указывая значения соответствующих грамматических, синтаксических и семантических параметров. Параметр «Часть речи спец.» содержит, помимо указания на часть речи, еще и определение семантической функции конструкта (единицы языкового реестра). Этот параметр используется как «заглавная» часть конструкта для составления моделей предложений в модуле «Синтаксис и фразеология».

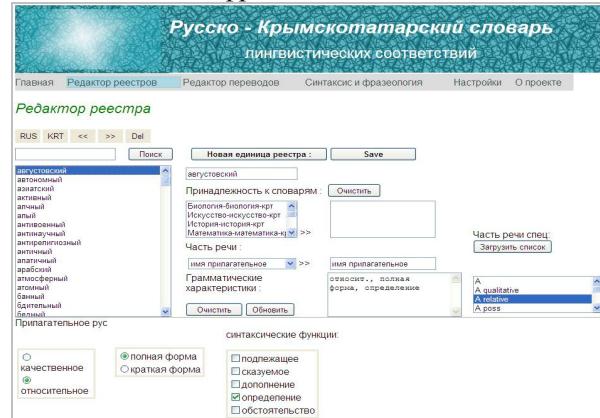


Рис.1. Редактор реестра.

Для представления в электронном словаре коллекции лингвистических соответствий синтаксических конструктов и фразеологизмов в составе инструментальной системы электронного словаря был разработан редактор «Синтаксис и фразеология» (Рис.2).

Входным языком для Реестра синтаксических конструктов и фразеологизмов может быть как русский язык, так и крымскотатарский. Каждая единица реестра помечается маркером принадлежности – «синтаксис» \leftrightarrow «фразеология» – и в дальнейшем обрабатывается разными программными модулями, подключенными к данному интерфейсу.

Обработка данных в модуле «Синтаксис» происходит следующим образом:

- из реестра выбирается пара синтаксических конструктов ИЛИ вводится новая пара + проставить флаг «Фразеология» + «Сохранить»;
- если схемы предложений уже составлены, они будут продемонстрированы в соответствующих текстовых полях, где их можно отредактировать.

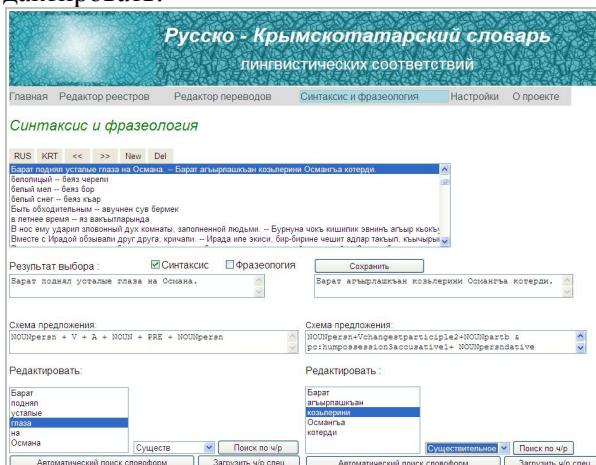


Рис.2. Синтаксис и фразеология. Процесс обработки данных в модуле «Синтаксис и фразеология».

Для составления схем предложений используются нижние группы элементов управления для каждого языка. Предложение разбивается на слова, осуществляется поиск соответствующей словоформы в главных языковых реестрах словаря и реестрах словоформ. Поиск возможен как с указанием части речи, так и без указания. Указание части речи должно сокращать время поиска. Если словоформа найдена, то ее параметры заносятся в схему предложения по шаблону: 1) аббревиатура части речи + семантическая функция, 2) грамматические параметры словоформы. Если поиск не дал результатов, в схему предложения можно внести соответствующую информацию вручную. Кроме составления схемы предложения, модуль «Синтаксис и Фразеология» производит индексирование синтаксического конструкта реестровыми индексами входящих в него концептов. Это приводит к тому, что в словарных статьях концептов накапливаются синтаксические примеры употребления данного концепта. Причем такой пример включает в себя синтаксические конструкты на двух языках (лингвистическое соответствие) и модели, что может успешно использоваться для обучения языку.

Подмодуль «Фразеология» также производит индексирование для демонстрации фразео-

логических соответствий в словарных статьях концептов главных языковых реестров. Инструментальная система электронного словаря разработана как веб-приложение (C#, Java), интегрированное с сервером баз данных SQL Server. На данном этапе инструментальная система и лексикографическая база данных словаря развернута в тестовом режиме в Научно-исследовательском центре крымскотатарского языка, литературы и истории имени Бекира Чобан-заде.

В настоящей статье впервые описаны категориальные признаки грамматически связанных и грамматически несвязанных конструктов крымскотатарского предложения, основные факторы их функционирования и образования, рассмотрены процессы взаимодействия лексической и грамматической семантики компонентов каждого конструкта в отдельности и в составе предложения, разработаны модели синтаксических конструктов с учетом параметров морфосинтаксической и лексико-семантической разметки, представлена фундаментальная теория крымскотатарского словосочетания, выявлены трансформационные возможности, коммуникативные, лексико-семантические, психолингвистические и прагматические особенности крымскотатарского предложения, состоящего из структурных и информационных, логических и фактических конструктов, разработаны параметры морфосинтаксической и лексико-семантической разметки. Кроме того, впервые с учетом параметров морфосинтаксической и лексико-семантической разметки разработан модуль «Синтаксис и Фразеология» для «Русско-крымскотатарского словаря лингвистических соответствий». Это позволяет в полуавтоматическом режиме создавать модели синтаксических конструктов на базе параллельных текстов.

Литература

1. Акмоллаев Э.С. Классификация бессоюзных сложных предложений и особенность их соотношения с типами союзных предложений (на материале бессоюзных сложных предложений открытой структуры): автореф. дис. ... канд. филол. наук: 10.02.08. Ташкент, ТГПИ, 1986. 53 с.
2. Рахилина Е.В. Лингвистика конструкций / Под ред. Ответственный редактор Е.В.Рахилиной, Т.И.Резниковой. М.: Азбуковник, 2010. 584 с.
3. Эмирова А.М. Крымскотатарская лексикография: современное состояние и перспективы развития // Культура народов Причерноморья: Научный журнал. Симферополь, 1997. №3.

- сентябрь. URL: <http://turkology.tk/library/161> (дата обращения 10.11.2014).
4. Щерба Л.В. Языковая система и речевая деятельность. Изд. 2-е, стереотипное. М.: Едиториал УРСС, 2004. С. 265-304; Герд А.С., Ивашико Л.А., Лутовинова И.С. и др. Основные типы словарей в отечественной русистике // Лексикография русского языка. Учебник для высших учебных заведений. СПб.: Факультет филологии и искусств СПбГУ, 2009. 672 с.
 5. Языки мира, находящиеся под угрозой исчезновения // Unesco.org/new/ru/unesco: ЮНЕСКО 2009-2014. URL: <http://www.unesco.org/new/ru/culture/themes/endangered-languages/atlas-of-languages-in-danger/> (дата обращения 10.11.2014).
 6. Виноградов В.В. Из истории изучения русского синтаксиса. М.: Изд-во Моск. ун-та, 1958. 400 с.; Грамматика русского языка / редкол.: акад. В.В. Виноградов, чл.-кор. АН СССР Е.С. Истрина. М.: Изд-во Акад. наук СССР, 1954. Т. 2: Синтаксис. Ч. 1. 703 с.
 7. Абдуллаев Э.М. Русско-крымскотатарский учебный словарь: более 5000 слов. Симферополь: Крымучпедгиз, 1994. 384 с.; Краткий словарь когнитивных терминов / Е.С. Кубрякова [и др.]. М.: Фил. фак. МГУ, 1997. 245 с.; Конструкт // Философский словарь. Библиотека «Полка букиниста». Значимые книги отечественных и зарубежных авторов. URL: <http://philosophy.polbu.ru/konstrukt.htm> и др. (дата обращения 10.11.2014).
 8. Меметов А.М. Земаневий къырымтатар тили. Симферополь: Къырым девлет оқыув педагогика нешрияты, 2006. 320 с. (Къырымтатар тилинде); Меметов А.М. Крымтатарский язык. Ч. 1. Общие сведения о языке; Ч. 2. Морфология : учеб. пособие. Симферополь: Крымучпедгиз, 2003. 287, [1] с.
 9. Бускунбаева Л.А., Сиразитдинов З.А. К системе разметок в национальном корпусе башкирского языка // Актуальные проблемы диалектологии языков народов России. Материалы XI межрегиональной конференции. Уфа. С. 50-55; Дыбо А.В., Шеймович А.В. Автоматический морфологический анализ для корпусов тюркских языков // Филология и культура. 2014. №2(36). С. 20-26; Захаров В.П. Корпусная лингвистика: учебно-методическое пособие. Санкт-Петербург: Санкт-Петербургский гос. университет, 2005. 48 с.; Корпусна лінгвістика: монографія / В.А. Широков [та ін.]; НАН України, Укр. мов.-інформ. фонд. Київ: Довіра, 2005. 472 с.; Корпусы по памятникам тюркских языков. URL: http://www.tuvancorpus.ru/?q=korporusy_po_pamyatnikam_tyurkskikh_yazykov и мн. др. (дата обращения 10.11.2014).
 10. Бектаев К.Б. Статистико-информационная типология тюркского текста : автореф. дис. на здобуття наук. ступеня д. фіол. наук: спец. 10.02.21: Структурная, прикладная и математическая лингвистика; Академия наук СССР, Институт языкоznания, Ленинградское отделение. Ленинград, 1975. 39 с.; Дыбо А.В., Шеймович А.В. Автоматический морфологический анализ для корпусов тюркских языков // Филология и культура. 2014. №2(36). С. 20-26; Захаров В.П. Корпусная лингвистика: учебно-методическое пособие. Санкт-Петербург: Санкт-Петербургский гос. университет, 2005. 48 с.; Национальный корпус русского языка. URL: <http://www.ruscorpora.ru>; Корпус по памятникам рунического письма Горного Алтая. URL: <http://www.altay.uni-frankfurt.de/>; Корпусна лінгвістика: монографія / В.А. Широков [та ін.]; НАН України, Укр. мов.-інформ. фонд. Київ: Довіра, 2005. 472 с.; Корпусы по памятникам тюркских языков. URL: http://www.tuvancorpus.ru/?q=korporusy_po_pamyatnikam_tyurkskikh_yazykov и мн. др. (дата обращения 10.11.2014).
 11. Дыбо А.В., Шеймович А.В. Автоматический морфологический анализ для корпусов тюркских языков // Филология и культура. 2014. №2(36). С. 20-26; Шеймович А.В. Морфологическая разметка корпуса хакасского языка // Российская тюркология. № 2(5). С. 48-61.
 12. Электронный корпус хакасского языка. URL: <http://khakas.altaica.ru/grammar/> (дата обращения 10.11.2014).
 13. Дыбо А.В., Шеймович А.В. Автоматический морфологический анализ для корпусов тюркских языков // Филология и культура. 2014. №2(36). С. 20-26.
 14. Крылов С.А. Стратегии применения интегрированной информационной среды StarLing в корпусной лингвистике и в компьютерной лексикографии // Orientalia et classica. Труды Института восточных культур и античности. Выпуск XIX. Аспекты компаративистики. З. М., РГГУ, 2008. С. 649-668.
 15. Корпус по памятникам рунического письма Горного Алтая. URL: <http://www.altay.uni-frankfurt.de/> (дата обращения 10.11.2014); Корпусы по памятникам тюркских языков. URL: http://www.tuvancorpus.ru/?q=korporusy_po_pamyatnikam_tyurkskikh_yazykov (дата обращения 10.11.2014); Электронный корпус по памятникам доисламских древнетюркских текстов. URL: <http://vatec2.fkidg1.uni-frankfurt.de/> (дата обращения 10.11.2014).

16. Электронный корпус шорского языка. URL: <http://shoriya.ngpi.rdtc.ru> (дата обращения 10.11.2014).
17. Машинный фонд башкирского языка. URL: www.mfbl.ru (дата обращения 10.11.2014).
18. Салчак А.Я. Электронный корпус тувинского языка / Международная научно-практическая конференция, посвященная 100-летию со дня рождения «Народного академика» Владимира Михайловича Наделяева. Кызыл // Новые исследования Тувы. URL: http://www.tuva.asia/journal/issue_15/5231-salchak.html (дата обращения 10.11.2014).
19. Бектаев К.Б. Статистико-информационная типология тюркского текста: автореф. дис. на заседании науч. ступеня д. филол. наук: спец. 10.02.21: Структурная, прикладная и математическая лингвистика. Академия наук СССР, Институт языкоznания, Ленинградское отделение. Ленинград, 1975. 39 с.
20. Портал государственного языка. URL: <http://til.gov.kz/wps/portal/> (дата обращения 10.11.2014).
21. Лингвистический корпус крымскотатарского языка. URL: <http://korpus.juls.savba.sk/QIRIM/#id9> (дата обращения 10.11.2014).
22. Оказ Л.С. Сопоставительная типология причастий крымскотатарского и русского языков // Ученые записки Таврического национального университета им. В.И.Вернадского. Серия «Филология. Социальные коммуникации». 2009. Т. 22(61), № 3. С. 46-50; Оказ Л.С. Типологически значимые аспекты сопоставления грамматических категорий крымскотатарского и русского языков (именные части речи) // Культура народов Причерноморья. 2008. №147. Т.2. С. 96-99.
23. Оказ Л.С. Средства выражения синтаксических отношений между частями сложного предложения в крымскотатарском языке // Культура народов Причерноморья. 2001. № 17. С. 184-185.

КЫРЫМТАТАР ТЕКСТЛАРЫНЫҢ КОРПУС АЛДЫ ЭШКӘРТМӘСЕ

Лемара Сергеевна Селендили,

М.А.Шолохов исемендәгә Мәскәү дәүләт гуманитар университеты,
Россия, 109240, Мәскәү ш., Верхняя Радищевская урамы, 16/18 нче йорт,
lemara2002@hotmail.com.

Мәкалә кырымтатар теле материалында синтаксик төзелмәләр теориясен яктыруға бағышланған. Тел материалының корпус алды эшкәртмәсе кырымтатар теленнән яки кырымтатар теленә машина тәржемәсенең тәңгәл системасы өчен теоретик база булдыру, телне һәм телләрне уқыту практикасын заманча мәғълүмати технологияләр ярдәмендә үстерүгә юнәлтелгән электрон сүзлекләр һәм инструменталь система оештыру максатыннан башкарыла.

Модельләрне төзу барышында, құләмле белешмәләргә таянып, текстның шартлары, автомат/ярымавтомат синтаксик тамгалар эзлеклелеге исәпкә алына. Синтаксик төзелмәләрне куллану даирәсө булып, «Лингвистик тәңгәллекләрнең русча – кырым татарча сүзлеге» лексикографик белешмә базасы өчен төзелгән «Синтаксис һәм фразеология» модуле тора.

Төп төшөнчәләр: кырымтатар теле, синтаксис, структур, гамәли һәм математик лингвистика, лингвистик корпус, текстның корпус алды эшкәртмәсе.