Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 95 (2013) 68 – 74

Procedia
Social and Behavioral Sciences

5th International Conference on Corpus Linguistics (CILC2013)

National Corpus of the Tatar Language "Tugan tel": Grammatical Annotation and Implementation

Dzhavdet Suleymanov[a,b], Olga Nevzorova[a,b,*], Ayrat Gatiatullin[a,b], Rinat Gilmullin[a,b], Bulat Khakimov[a,b]

[a]*Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Bauman str., 20, Kazan, 420111, Russia*
[b]*Kazan (Volga Region) Federal University, Kremlevskaja str., 18, Kazan, 420008, Russia*

Abstract

This article presents the National Corpus of the Tatar Language, which is being developed at the Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences on the EANC technological platform. It describes the morphological model of the Tatar language used for grammatical annotation of words.

Keywords: Tatar language; Turkic languages; morphological annotation

1. Introduction

Nowadays the projects of developing generally accessible electronic corpora of Turkic languages are topical. Among well-known projects are the corpora of Turkish (Aksan, Y. et al.,2012; Dalkiliç, G., & Çebi Y., 2002; Say, Bilge, Deniz Zeyrek, Kemal Oflazer & Umut Özge, 2002), Uighur (Yusup Aibaidulla, & Kim-Teng Lua, 2003), Bashkir (Buskunbaeva L.A., & Sirazitdinov Z.A., 2011), Hakass (Sheimovich, 2011), Kazakh (http://til.gov.kz) and Tuvan (Salchak, 2012) languages. They are currently on different phases of project realization and most of them are monolingual. The Uyghur-Chinese corpus is an example of a parallel corpus. Different software is being developed on its basis; in particular, the corpus is used for designing systems of statistical machine translation. According to

* Corresponding author. Tel.: +7-905-022-0318: fax: +7-843-292-6888
E-mail address: onevzoro@gmail.com

doi:10.1016/j.sbspro.2013.10.623

the level of completeness of the annotation system, the most successful are the electronic corpora of the Turkish and Uighur languages. They have implemented annotation systems for different language levels. Other above-mentioned projects of electronic corpora of Turkic languages are in the initial phase of their development. Therefore, it can be stated that the Turkic corpus linguistics is currently in the process of corpora formation for the languages of the Turkic group.

The development of the electronic corpus of the Tatar language is based upon early works related to the creation of the culmination of electronic resources of the Tatar language (Bukharaev, 1995). Nowadays, scientists of the Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences carry out the project of designing the National Corpus of the Tatar Language "Tugan Tel". It is done within the Fundamental Research Program of the Russian Academy of Sciences "Corpus Linguistics" 2012-2014. The National Corpus of the Tatar Language is developed as a monolingual corpus with grammatical annotation of words; semantic annotation will be added to it in the nearest future. The corpus is implemented on the EANC technological platform (http://www.eanc.net/).

This article describes the composition of the corpus, the morphological model which is used for grammatical annotation, as well as the issues of current implementation of the Tatar corpus.

2. Composition of National Corpus of Tatar Language

The National Corpus of the Tatar Language can be viewed as a set of conceptual and functional models of different levels of the Tatar language (Suleymanov D.Sh. & Gatiatullin A.R., 2003). The class of conceptual and functional models includes structural and functional descriptions of a certain linguistic level (or levels), as well as different types of general information, which is necessary for developing of information systems and technologies of natural language processing.

The corpus is an open system, therefore it permits the expansion of the annotation system (currently only grammatical annotation is used). The Tatar Corpus contains texts of different genres and styles of the modern literary Tatar language. The main sources of electronic copies of texts for the corpus are fictional texts, educational and scientific literature, texts of Internet publications on informative, social and political themes and texts of official documents. In the future we plan to reinforce the chronological and genre balance of the corpus, i.e. through digitalization of printed texts of the Soviet period.

The distribution of texts in the National Corpus of the Tatar Corpus according to their genre is presented in Table 1. In the current version of the corpus the texts are divided into two types: fiction and non-fiction. In the future a more detailed classification of genres of texts will be introduced.

Table 1. Distribution of texts according to their genre in the National Corpus of the Tatar Language

Genre	Amount of words	Share in the corpus, %
Fiction	19 279 033	71,45 %
Non-fiction	7 703 258	28,55 %
Total	26 982 291	100 %

A set of metadata is related to each input text. It reflects the structure of the database where the collection documents are stored. To present a textual document, the following set of basic descriptors is employed:

- Document number
- File name
- Type of text (original/ translation)
- Language of the document
- Name of the work
- Author of the work
- Last name of the translator (for a translated document)
- Genre (fiction/ non-fiction)
- Year of edition
- Number of words in the document (word usages)

- Existence/ non-existence of translation into Russian
- Number of words in the text in Russian
- Source of the original document
- Note field for the original document
- Source of translation (editorial)
- Note field for the translated document
- Check-up of the document by a corrector

3. Morphological model and grammatical annotation

The system of morphological annotation of the National Corpus of the Tatar Language is mainly oriented in presenting of all the existing grammatical word-forms, which are not always reflected in the descriptive researches on Tatar Grammar or have different alternative interpretations. In the model used for formal representation of the Tatar agglutinative morphology a word-form is built upon consecutive adding to the base of regular word-formative and inflectional affixes. As a rule, each grammatical meaning is expressed by a separate affix, and the affixes are unambiguous and regular. Thereby, in order to mark up a word, it is necessary to analyze the structure of its affixal chain, sometimes making use of the stems dictionary.

Grammatical annotation of a Tatar word includes the information about the part of speech of the word and a set of morphological features (parameters). Taking into consideration the characteristics of the Tatar morphotactics, grammatical parameters are divided into the complex / simple, from the one hand, and compulsory /optional, from the other hand. All the complex parameters are represented by a set of affixes associated to a grammatical category (for example, the case category has a set of case affixes), while the simple ones are represented by a single affix (for example, the category of interrogation has a single interrogative form, as can be seen in the example (1)). A compulsory parameter is always present in the description of a word-form in a certain part of speech (nouns are always placed in a certain case, there are no “no-case” nouns). In the case of an optional parameter, the affix which expresses the grammatical meaning is facultative (i.e. the meaning of possessiveness is not necessarily expressed in the Tatar nouns by affixal means).

Let us consider the examples (2), (3) and (4). The examples (2) and (3) are the examples for explicit representation of case and possession, respectively. In the example (4), although the word “yort” doesn’t have any case affix, case as an always-required (obligatory) parameter is represented implicitly. But possession as a facultative (optional) parameter is not represented in the example (4) at all.

Note that all the examples have the following structure:

- line 1: the word-form(s) in Tatar (Cyrillic alphabet)
- line 2: the word-form(s) in Tatar presented as a stem followed by a sequence of affixes (Latin alphabet)
- line 3: translation into English of the stem of each word-form in Tatar with grammatical tags
- line 4: translation of the word-form(s) in Tatar into English

(1) Син Татарстаннанмы?
Sin Tatarstan-nan-mı?
PRO N+ABL+INT
Are you from Tatarstan?

(2) йортта
yort-ta
N+LOC
in the house

(3) йортым
yort-ım
N+POSS.1SG
my house

(4) йорт
yort-0
N.NOM
a house

While working on the annotation system, a certain morphological standard of the corpus has been developed. It contains the ways of interpretation of some morphological phenomena of the Tatar language, especially the debatable ones. In our opinion, it is necessary to keep the balance between the objective reality and traditions of grammatical theory. Therefore, for instance, multifunctional affixes with doubtful status are included into different paradigms of the morphological model of the corpus. They are enumerated in the list of grammatical features which is available when we create a search query.

In examples (5) - (7) we can see such polyfunctional affixes: -lı/-le and -sız/-sez. These are attributivizers, which are usually described in Tatar grammar books as derivative suffixes producing adjectives. Alongside with this, they also may express syntactical relations of presence or absence: munitive and abessive. In example (5) *köçle* is an adjective derived from the noun *köç (strength)*, and -le is not separated. In examples (6) and (7) *tübä-le* and *aqça-sız* are not adjectives, but inflected nouns meaning '*with roof*' and '*without money*', respectively. These characteristics are reflected in the annotation of our corpus.

(5) Урамда көчле яңгыр ява
uram-da köçle yaňğır yaw-a
N+LOC ADJ N V+PRES
It's heavy raining in the street

(6) Яшел түбәле йорт
yäşel tübä-le yort
ADJ N+ATTR.MUN N
a house with green roof

(7) Ул акчасыз килде
ul aqça-sız kil-de
PRO N+ATTR.ABES V-PST
he/she came without money

Morphological annotating of corpus texts is carried out using the module of two-level morphological analysis of the Tatar language implemented in the program tool PC-KIMMO (http://www.sil.org/pckimmo/about_pc-kimmo.html). The list of grammatical tags used in the process of morphological annotation of the corpus is located in the file of morphotactic rules of the software module. For complete description of the morphological model of the literary Tatar language about 60 morphological tags are used. A list of part-of-speech tags is given in Table 2.

Table 2. Part-of-speech tags in the Tatar corpus

Part of speech (English)	Part of speech (Tatar)	Tag
Noun	İsem	N
Adjective	Sıyfat	ADJ
Verb	Fiğel	V
Adverb	Räweş	ADV
Numeral	San	NUM
Pronoun	Almaşlıq	PRO
Conjunction	Terkägeç	CNJ
Postposition	Bäylek	PART
Particle	Kisäkçä	POST
Interjection	Imlıq	INTJ

Modal word	Modal süz	MOD
Imitative word	İyärtem	IMIT

To denote morphological categories expressed by corresponding morphemes, a system of denotations has been worked out. It takes into consideration modern general typological and Turkic researches (Mishar dialect, 2007) and corresponds to the generally accepted international terminology (Leipzig glossing rules, http://www.eva.mpg.de/lingua/resources/glossing-rules.php). In Table 3 you can see some tags developed particularly for the Tatar language.

Table 3. Specific grammatical tags in the Tatar corpus

Category	Affix	Tag
attributive munitive	-lı/-le	ATTR.MUN
attributive abessive	-sız/-sez	ATTR.ABES
attributive locative	-dağı/-däge	ATTR_LOC
attributive genitive	-nıqı/-neke	ATTR_GEN

Let us consider the examples of grammatical annotation of words belonging to different parts of speech. Let us assume that the following sentence is put into the analyzer:

(8)	Без	урамыбыздан	үтеп	барабыз
	Bez	uram-ıbız-dan	üt-ep	bar-a-bız
	N / PRO.1PL	N+POSS.1PL+ABL	V+CONV	V+PRES+1PL
	We are going along our street			

Each word-form in the example (8) receives automatic morphological analysis, recorded in a separate line of the file with the analyzed text. At the moment the grammatical homonymy is not removed in the corpus, so alternative analysis is presented for homonymic word-forms in the process of annotation. Thus, for example, “bez” is homonymic, and two alternative mark-up options (N(bez) / PRO.1PL(bez)) are offered.

We need to mention that different infringements of regularity of the Tatar language morphology lead to difficulties in automatic processing, because many morphotactic rules do not work on this material. Many of these infringements are caused by a big number of non-assimilated borrowings and not perfect modern Tatar orthography.

4. Implementation of the corpus on the EANC platform

The National Corpus of the Tatar Language is put on the Internet using the EANC platform, originally developed by the company *CorpusTechnologies* for the East Armenian National Corpus in 2007. The platform consists of the search system, web-interface and indexer – a module, which transforms input text files into databases and files used by the search system. Although the platform was initially used for the Armenian language, the majority of its functional possibilities is universal and can be used for presenting corpora of texts in any languages. In particular, in 2011 within the Program of the Presidium of the Russian Academy of Sciences “Corpus Linguistics” on the base of EANC platform the corpora of the Albanian, Kalmyk, Lezgian and Ossetian languages were created.

Figure 1 presents the overall look of the main page of the National Corpus of the Tatar Language on EANC platform.

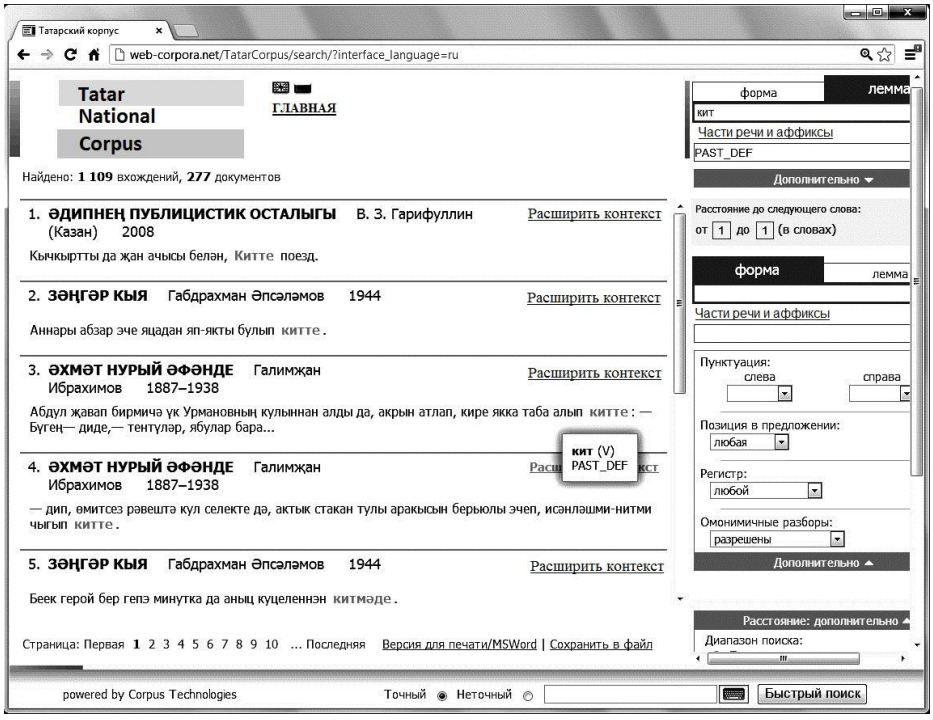

Fig. 1. Main page of the National Corpus of the Tatar Language

The input Tatar texts with morphological tagging, when downloaded into the system, are converted into XML-format. Below there are XML-presentations of words from the example (8):

<w><ana lex="bez" gr="N"; lex="bez" gr="Pro_1Pl"></ana>bez</w>

<w><ana lex="uram" gr="N, Poss_1Pl, Abl"></ana>uramızdan</w>

<w><ana lex="üt" gr="V, Conv"></ana>ütep</w>

<w><ana lex="bar" gr="V, Pres, 1Pl"></ana>barabız</w>

The word-homonym is encoded by all possible sets of grammatical features. In the current realization of the Tatar corpus the homonymy is not removed (see the encoding of the word “bez” in the example).

The search system supports the following types of user queries: search by the exact form, by lemma (initial form) and by a number of grammatical parameters, as well as a combination of these queries. As result of the query processing, the sentences which contain the words corresponding to established criteria are selected. The results of each query are shown on the output page (number of sentences on each page can be chosen from the settings and cannot exceed 50).

5. Conclusion

The article describes the experimental version of the National Corpus of the Tatar language “Tugan Tel” 1.0. The current amount of words in the corpus is 20 million. The corpus is now on the stage of data testing. It includes mainly prosaic texts which represent the literary Tatar language (from the 20th of the XIX century in Cyrillic alphabet), as well as modern scientific and business texts, texts of official documents and newspaper materials. All the texts included in the Tatar corpus go through special procedures of meta-annotation (attributing of metadata to the text) and morphological annotation (attributing of morphological information to each word-form).

The developed morphological model and system of grammatical annotation present an attempt of complex realization of detailed, theoretically founded and pragmatically oriented grammatical annotation in the corpora of Turkic languages. The main problems when working on this task are the absence of a universal language of

grammatical categories' description, polysemy and polyfunctionality of morphemes, difficulty in separating of "pure" meanings. The offered system of grammatical annotation for the corpus of the Tatar language considers all these aspects.

Acknowledgements

This work was supported by the Fundamental Research Program of the Russian Academy of Sciences "Corpus Linguistics" 2012-2014.

References

Aksan, Y. et al. (2012). Construction of the Turkish National Corpus (TNC). *Proceedings of the Eight International Conference on Language Resources and Evaluation* (LREC 2012). İstanbul, Turkiye. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/papers.html

Dalkiliç, G., & Çebi Y. (2002). A 300 MB Turkish Corpus and Word Analysis. Advances in Information Systems, 2d International Conference, ADVIS 2002, Izmir, Turkey. Proceedings. LNCS 2457, 205-212.

Say, Bilge, Deniz Zeyrek, Kemal Oflazer and Umut Özge (2002). Development of a Corpus and a TreeBank for Present-day Written Turkish. *Proceedings of the Eleventh International Conference of Turkish Linguistics*, Eastern Mediterranean University, Cyprus, August 2002.

Yusup Aibaidulla, & Kim-Teng Lua (2003). The Development of Tagged Uyghur Corpus. *Proceedings of PACLIC17*, 1-3 October 2003, Sentosa, Singapore, 228-234.

Buskunbaeva L.A., & Sirazitdinov Z.A. (2011). The System of Annotation in the National Corpus of Bashkir Language. *Proceedings of the International Conference "Languages of Minorities in Computer Technologies: Experience, Tasks and Perspectives"*, Yoshkar-Ola, 46-51. In Russian.

Sheimovich A.V. (2011). Morphological Annotation of the Corpus of the Hakass Language. *Russian Turkology*, 2(5), 48-61. In Russian.

Salchak A.Ya. (2012). Electronic Corpus of the Tuvan Language. Electronic informational magazine "New Researches in Tuva", 3. Retrieved from URL: http://www.tuva.asia/journal.

Bukharaev R.G., & Suleymanov D.Sh. (1995). To the Conception of Implementation of the Tatar Language into Computer Technologies. *Tatar Language and New Information Technologies*, Issue 2, Kazan: Kazan State University, 8-19. In Russian.

Suleymanov D.Sh., Khakimov B.E., & Gilmullin R.A. (2011). Corpus of the Tatar Language: Conceptual and Linguistic Aspects. *TGGPU Digest*, 26, 211-216. In Russian.

Suleymanov D.Sh., & Gatiatullin A.R. (2003). *Structural and Functional Model of Tatar morphemes*, Kazan: Fen. In Russian.

Mishar Dialect (2007). *The Mishar Dialect of the Tatar Language: Essays on Syntax and Semantics*. Eds. E.A. Lutikova, K.I. Kazenin, V.D. Solovyev, & S.G. Tatevosov, Kazan: Magarif. In Russian.