EPiC Series in Language and Linguistics
Volume 1, 2016, Pages 121–129
CILC2016. 8th International
Conference on Corpus Linguistics

On the Way to the Relevant Grammatical Tagset for the Tatar National Corpus

Alfia Galieva[1], Bulat Khakimov[1,2] and Airat Gatiatullin[1]
[1] Applied Semiotics Research Institute of the Tatarstan Academy of Sciences
[2] Kazan Federal University, Russia.
amgalieva@gmail.com, bulat.khakeem@gmail.com, agat1972@mail.ru

Abstract

The development of the metalanguage for annotation is one of the topical issues in modern corpus linguistics. One of the main problems in the development of a grammatical tagset for the Tatar National Corpus is to identify the inventory level of inflectional categories and to create an optimal metalanguage of description. We discuss the factors that complicate the process of grammatical annotation for Turkic corpora in general, including the need to overcome the influence of the Indo-European grammatical tradition in the description of the phenomena of Turkic languages, the lack of generally accepted standards for corpus annotation, the lack of a common metalanguage used to describe grammatical categories of Turkic languages, poor differentiation of word-building and form-building in Turkic languages, etc. In the course of work on the system of grammatical annotation of the Tatar Corpus, we made an inventory of grammatical categories of the Tatar language and developed a metalanguage for describing them.

Currently, the developed grammatical tagset contains 93 tags. Tags for parts of speech and grammatical categories were created to meet the worldwide standards, primarily the Leipzig glossing rules.

1 Introduction

Development of Turkic studies during the past years has been marked by deepening the theoretical foundation of a linguistic research, with an increasing emphasis on new directions and challenges of modern linguistics, including applied linguistics. Criteria and principles of analysing grammatical categories remain one of the priority directions of research in Turkic languages, and nowadays the interest to this traditional subject is supported by formulating new tasks, posed by the development of information technologies. In recent years, a number of Turkic corpora has been developed, and these corpora, together with a new interest in empirical linguistic data and validation of theory and linguistic description make for a surge of novel work using corpus methods to study the grammar of

A.Moreno Ortiz and C.Pérez-Hernández (eds.), CILC2016 (EPiC Series in Language and Linguistics, vol. 1), pp. 121–129

Turkic languages.

The Tatar language belongs to the Turkic group that forms a subfamily of Altaic languages. The Tatar language is spoken in West-central Russia (in the Volga region) and southern parts of Siberia. The number of Tatars in Russia in 2010 was 5.31 million people.

The results of theoretical studies on grammar are used for corpus development and annotation. The Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences, together with Kazan Federal University, continue the work on "Tugan Tel" Tatar National Corpus (http://corpus.antat.ru). This corpus is built as a monolingual grammatically annotated corpus. The Tatar National Corpus contains texts of different styles and genres of modern Tatar literary language (fiction, media texts, official documents, educational and scientific literature, etc.).

The Tatar National Corpus project is carried out within the framework of the State Program named "Preservation, study and development of the official languages of the Republic of Tatarstan and other languages in the Republic of Tatarstan for 2014-2020". The developed Corpus is intended for a wide range of users: for linguists, specialists in the Tatar language and culture, teachers of the Tatar language, cultural workers, and for everyone who is interested in studying the Tatar language. The volume of the Corpus had reached 82,000,000 tokens by the end of 2015.

The Tatar National Corpus has a system of grammatical annotation that is oriented at presenting all the existing grammatical word-forms. Grammatical annotation of a Tatar word includes the information about the part of speech of the word and a set of morphological features. Morphological annotation is carried out using our own morphological analyzing tool which was created on the basis of the PC-KIMMO two-level morphology model [2013]. The search functionality of the Corpus includes search queries for lemmas (lexemes), word forms, and individual grammatical features.

This paper is organized as following: part 2 outlines related works, part 3 gives general information on typological features of Tatar and formal representation of Tatar agglutinative morphology; part 4 describes the main challenges that occurred in the development of the system of grammatical annotation, and proposes solutions to the problems encountered.

2 Related works

Nowadays, projects of developing electronic corpora of Turkic languages are quite relevant. Among the well-known projects of electronic corpora of Turkic languages we can mention the electronic corpora of Turkish [Aksan, Y. et al, 2012; Dalkiliç, G. and Çebi Y., 2002; Say et al, 2002], Uighur [Yusup Aibaidulla and Kim-Teng Lua, 2002], Bashkir [Buskunbaeva L.A., 2011], Khakassian [Sheimovich, 2011], Kazakh [http://til.gov.kz] and Tuvan [Salchak 2012] languages. These corpora are at different stages of implementation and are mostly monolingual. However, a look at the systems of grammatical annotation in these corpora reveals a variety of different theoretical approaches and empirical foci which can be traced back to different linguistic traditions and research paradigms.

[Dybo & Sheymovich 2014] describe the main principles on which the automatic morphological analyzer for Turkic corpora operates. The authors make an inventory of main components of the automatic morphological analysis system: a grammatical dictionary; a range model of the word form (including a set of ranges with a series of morpho-phonological forms of inflectional affixes for each range); a set of compatibility rules for affixes and a two-level set of phonetic rules that constrain the choice of components of the word form.

The paper [Galieva, Khakimov & Gatiatullin, 2013] discusses some general issues of reflecting grammatical information in the Tatar National Corpus. Creating a corpus tagset goes beyond a purely applied problem, inevitably making it necessary to solve numerous theoretical problems which, for many years, have had alternative interpretations in Tatar linguistics depending on research goals and aspects. It is determined that various problems that arise when creating a metalanguage for the

description of the structure of Tatar word forms are due to the lack of standards on the development of corpus annotations and universal terminology, ambiguity and homonymy of affixes, the requirement of compatibility with other electronic linguistic resources, etc.

3 Formal representation of Tatar agglutinative morphology

The most important phonetic feature of Turkic languages is progressive vowel harmony. The basic way of word formation and inflection is progressive affixal agglutination when a new unit is built by consecutive addition of regular and clear-cut monosyllabic derivational and inflectional affixes to the stem, therefore the stem remains unchanged. Affixal agglutination provides unified morphological means for forming derivatives within the same grammatical class of words as well as for changing the part-of-speech characteristic of the word and for turning it into another lexical or grammatical class. The boundaries between the affixes within the word form are distinct and transparent, and the affixal joint in many cases coincides with the syllabication [Guzev, 1981].

In Turkic languages the order of affixes is rigidly determined, and derivational affixes (e.g. suffixes) precede inflectional ones. Each added suffix tends to modify the whole preceding stem. Words have no classifying categories, like grammatical gender or animacy. Most affixes in the affix chain are unambiguous. There is one type of declension and conjugation in which one set of affixes is used only. The Tatar language has no grammatical prefixes and prepositions, although it has postpositions.

Another typologically relevant feature of the Tatar language is absence of clear-cut borders between inflection and derivation, since the same affixes in different positions may function both as inflectional and derivational.

Tatar nouns are marked with regard to their number, possession and case and are not characterized by definiteness. The Tatar verb has no aspect, but is characterized by tense, mood and can have the negative form.

The plural of nouns is formed by joining the affix -*LA*ʀ to the stem; the same plural affix is used to nominalize adjectives and to form 3d person of verbs. Possessive affixes are used to express the person and the number of possessors.

An example of the word formation and inflection below represents some salient features of Tatar morphology:

(1) *Qadaq-la-š-qan-nar-ı-na*
Qadaq -a nail, a tack (NOUN)
Qadaq-la – to nail (VERB)
Qadaq-la-š – to help to nail (VERB)
Qadaq-la-š-qan –[he, she] helped to nail (VERB, PARTICIPLE)
Qadaq-la-š-qan-nar- [they] helped to nail (VERB, PARTICIPLE)
Qadaq-la-š-qan-nar-ı – those who helped to nail (NOUN)
Qadaq-la-š-qan-nar-ı-n – to those who helped to nail (NOUN)

Tatar morphology is regular and predictable in many respects, and there is little or no fusion between the stem and the affixes.

For formal representation of Tatar agglutinative morphology, we use the model where the word form is built by consecutive addition of regular inflectional affixes to the stem. For example, a noun word form has the following regular structure: <stem> <plurality> <possessivity> <case> <modality>:

(2) *kitap-lar-ı-nan-mı*
book-PL, POSS_3SG, ABL, INT

'whether from his books', 'from his books?'

So a Tatar agglutinative word form is built by adding standard, mostly unambiguous, affixes to the stem, with the order of affixes and phonetic changes of affixes rigidly determined, and with affix boundaries clearcut. Nevertheless an attempt to build a paradigm of an individual word shows that this paradigm is extremely complicated and divaricate, consisting of a great number of inflectional affixes.

As a rule, every grammatical meaning is expressed by a particular affix. Affixes on the whole are regular and unambiguous. Thus, in order to tag a word form, it is necessary to analyze the structure of the sequence of its affixes, in some cases involving the dictionary of stems.

4 Challenges: the problem of relevance

One of the major problems in the development of grammatical annotation for the Corpus of the Tatar language is to identify the inventory level of inflectional categories in the Tatar language and to create an optimal metalanguage for description of these grammatical categories, which would be suitable for a wide range of potential users of the Corpus (specialists in Turkic studies, typologists, and lay users).

Development of the system of grammatical annotation for the Corpus has become a challenge for us. On the one hand, we relied on the information provided in academic Tatar language grammars [Tatar grammar 1993; Tatar grammar 2002]; on the other hand, specialized studies on general morphology and linguistic typology were involved [Plungian 2003 et al.]. Part-of-speech and grammatical categories tags are worked out taking into consideration the standards formed in the world, primarily the Leipzig Glossing Rules [Leipzig]. Grammatical annotation systems in existing corpora of other languages, including the Turkic ones, were also studied.

Here are some of the factors that hinder the process of grammatical annotation of corpora for Turkic languages in general:

- *Lack of generally accepted standards for corpus annotation*

A recognition of the importance of morpheme-by-morpheme glossing and presenting precise information about grammatical properties of individual words and parts of words is shared by many linguists and developers of corpora for morphologically rich languages. Nevertheless a standard way of presenting linguistic information has not been worked out yet. Different ways of representation of linguistic information are called forth by objective scientific problems (for example, great variety of languages and lack of transparency of morphological processes) as well as absence of an organizing and coordinating center [Kibrik 2004]. Specialists in linguistic typology mostly rely upon and widely use Leipzig Glossing Rules that provide certain standard ways of abbreviating possible descriptions.

Linguistic corpora can be provided with systems of annotation of a different nature. The most common form of grammatical annotation is when a word class label (part-of-speech tag) is assigned to words. This kind of annotation is implemented, for example, in the Brown Corpus, the LOB Corpus and the British National Corpus (BNC). Researchers have not agreed on a standard lexical and grammatical annotation model for English. A comparative evaluation of modern English corpus grammatical annotation schemes is presented in [Atwell et al., 2000].

There are special studies on a common metalanguage and tagset, for example, for Slavic languages [Derzhanski 2009, Sharoff 2008]. In 2014, on the Uniturk workshop which was held in Kazan, this problem was discussed for the first time for Turkic languages of Russia [Galieva et al, 2013].

- *Lack of a common metalanguage to describe grammatical categories of Turkic languages.*Turkology requires categorical means and a metalanguage that would gove an optimal

reflection of the grammatical structure and semantic system of Turkic languages. The organization of grammatical categories, forms and their meanings is strictly individual for each language. Therefore, grammatical and semantic annotation should reflect the uniqueness of the language system of particularly the Tatar language and other Turkic languages, and not blindly copy the concepts which were developed in the study of some other language and were assigned to the corresponding term.

A traditional linguistic categorical apparatus (given in grammars of Turkic languages) initially was created to describe the grammatical structure of Indo-European languages, and it is not always suitable to the categories of non-Indo-European languages. So we have to overcome the influence of foreign languages.

For example, the *Genitive case* in Latin, Greek and Russian is essentially the same grammatical category, whereas the so called Genitive in Tatar or in any Turkic language is quite a different thing. Nevertheless Turkic grammars use the term Genitive.

The *Reciprocal voice* in Turkic languages is a category which is quite different from the one with the same name in the Indo-European languages. The list can be easily continued.

Let us consider some typical difficulties in choosing the name and the tag for a grammatical category on an individual example. Working on the grammatical tagset for the Tatar National Corpus, we had great controversy regarding the annotation of Tatar adverbial-participial forms derived from verbs. There is a considerable diversity of these forms in Tatar and they are in active use.

Among the proposed variants were the following: CONV – converb, ADVV – adverbial verb, GER – gerund.

Each option has its advantages and disadvantages. The term “converb” is well-known to typologists, but it is barely used by Tatar linguists, so it is unfamiliar to a substantial part of corpus users, including specialists in the Tatar language. The term "converb' is used in the grammatical annotation system of the corpora of Minority Turkic languages.

The term “gerund” (GER tag is used in Bashkir corpus, and the Bashkir language morphologically is the closest to the Tatar language) is familiar to ordinary corpus users through the foreign languages they studied, but it does not correspond substantially to the respective category in the Tatar language, because it is used in the grammars of European languages and designates a specific verbal form of these languages. The expression “adverbial verb” is the translation of the Russian word *deeprichastiye* into English, and in its essence it is a kind of a calque for the Tatar term *häl fiğıl* used in Tatar grammars, but it is quite cumbersome for the corpus annotation. The term“adverbial verb” is used in the current version of the Tatar National Corpus annotation, but we are considering the reasonability of replacing it by the “*converb*” due to the increasing use of this term in modern works. It can be mentioned also that the CONV tag have been used in one of the earlier annotation systems of our corpus.

So in the course of work on the system of grammatical annotation of the Tatar National Corpus, we made an attempt to create a metalanguage of grammatical categories of the modern Tatar language, taking into consideration the Tatar linguistic tradition, as well as researches carried out within the framework of general Turkic theoretical studies and the achievements of modern linguistic typology.

- *Poor differentiation of word-building and form-building in Turkic languages, and a lack of clear boundaries between them.*

One of the consequences of this is the ongoing debate about the number of grammatical cases in Turkic languages. For example, the actual Tatar Grammars say that the Tatar language has 6 grammatical cases and some case-like forms. Table 1 represents a set of case tags.

Tag	Description	Affix
NOM	Nominative	-

GEN	Genitive	nIň
DIR	Directive	GA
DIR_LIM	Directive with limitative meaning	GAçA
ACC	Accusative	NI
ABL	Ablative	DAn
LOC	Locative	DA

Table 1. Tags for case affixes.

The Tatar grammar books note the so-called 'multifunctional' affixes which may express grammatical relations serving to link the words in a sentence, as well as to coin new words. One and the same affix can express grammatical meaning as well as derivational meaning (often in combination with the same stems). Ways of interpretation depend on the context. For example, a word *atnalık* in (3) is a noun, and in (4) it is an attributive form:

(3) *kitap atnalıgı uzdı*
book week-NMLZ, POSS-3 pass-
Book Week was held

(4) atnalık azık
week-NMLZ, food
some food for a week

The question is in what way we are to chose to mark such ambiguous affixes and if the tag must cover all the basic meanings.

- *Syncretism of grammatical categories.*

There is a small number of 'pure' grammatical categories in Tatar. The morpheme simultaneously expresses multiple grammatical meanings.

Let us take for example the meanings of two cases expressing the meaning of definiteness – the Accusative and the Genitive. The Accusative case expresses the meaning of the direct object and the definite object:

(5) hat yaza
letter_NOM wtite-PRES_3
writes *a* letter

(6) hatnı yaza
letter_ACC wtite-PRES_3
writes *the* letter [letter has been already mentioned]

What name for the category should be taken? The Accusative case, the Definite-Accusative case or any other?

The Genitive case also expresses at the same time the attributive meaning and the idea of definiteness.

(7) Шәһәр паркы
city park

(8) Шәһәрнең паркы
the park of the city [the referential use of the noun opposed to the non-referential use]

In the same way, verb tenses express the additional meaning of modality. The forms of the Past Tense – the meaning of the past and the meaning of subjective modality (evidentiality):

(9) Кичә яңгыр яуды
Yesterday it was raining [and I saw it].

(10) Кичә яңгыр яуган
Yesterday it was raining [but I did not see it].

The forms of the Future Tense additionally can express modality of epistemic possibility, degree of confidence (certitude) of the speaker in the commission of the act:

(11) Иртәгә яңгыр явар
[I think that] it will be raining tomorrow

(12) Иртәгә яңгыр явачак
It will be [definitely] raining tomorrow [and the speaker can't doubt it].

What name for the category should be taken? Past Definite, Past Indefinite, or Past Evidential?

- *Missing terms and descriptions*

Some grammatical forms in Tatar grammars do not have specific terms. For example, there is a group of attributive affixes which in grammar books are usually referred to in a descriptive way: '*form on –lı*', '*form on –sız*', etc.

For such forms, we have developed a tagging system based on the international terminology. Table 2 represents a set of tags for attributive forms derived from nouns.

Abbreviation/Tag	Interpretation of abbreviation	Affix	Example of word form	Structure of word form	English translation of word form
ATTR_MUN	attributive munitative	-lı	Maşina-lı	N+ATTR_MUN	Having a car
ATTR_ABES	attributive abessive	-sız	Maşina-sız	N+ATTR_ABES	Not having a car (carless)
ATTR_LOC	attributive locative	-dagı	Maşina-dagı	N+ATTR_LOC	That is in the car
ATTR_GEN	attributive genitive	-nıkı	Maşina-nıkı	N+ ATTR_GEN	That is of a car

Table 2. Attributive forms.

Here are the examples:

(13) mašina
car

(14) mašina-lı (keše)
(person) with a car, who has a car

(15) mašina-sız (keše)
(person) without a car, who does not have a car.

So in course of the work on the system of grammatical annotation of the Tatar Corpus, we made an inventory of grammatical categories of the Tatar language and developed a metalanguage for describing its grammatical categories. The work was carried out on the basis of Tatar grammar books taking into consideration the existence of alternative viewpoints on many issues. We also consulted research works on Turkic studies, linguistic typology and corpus linguistics. Part-of-speech tags and tags for grammatical categories were created to meet the standards formed in the world, primarily the Leipzig Glossing Rules. The metalanguage of grammatical tags in the Tatar National Corpus relies on the system of abbreviations based on the English language, due to the assumingly widespread international audience of the Corpus. In grammatical annotation the part of speech and inflectional features of the word are indicated as, for example, the category of possessiveness and case for nouns, ways of verbal action (grammatical raritives), aspect (negative form) or tense for the verb.

5 Conclusion

Currently, the developed system of grammatical annotation for the Tatar National Corpus contains 93 tags for words of different parts of speech. The work in this field is going on.

The work carried out on the corpus annotation of grammatical categories of the Tatar language indicates a lack of elaboration of many theoretical issues in Tatar linguistics and a need for amendments in grammar books of the Tatar language. These should be made taking into consideration the array of corpus data, which would provide objective information on the frequency and distribution of grammatical forms. A further research might be needed to develop common standards for data representation and description of the language material in the Turkic corpus linguistics. This will allow to elevate the comparative studies to higher modern standards and to create effective systems of automatic text processing for kindred languages.

In addition to deepening our knowledge and understanding of individual languages, corpus-oriented work on grammar has wider implications that concern methodological as well as theoretical aspects. Relevant topics and research questions concern, for instance, annotation schemata for larger syntactic units and syntactic relations, the increased use of advanced statistical methods and models in linguistics, the relation and boundary between grammar and discourse, and more generally the interface between corpus linguistics and linguistic theory.

References

Guzev, V., Nasilov, D. (1981). *Inflectional categories in the Turkic Languages and the Concept of "Grammatical Category"*. In: Soviet Turkology, 1981. N 3. pp. 22–35 (In Russian).

Suleymanov, D., Nevzorova, O., Gatiatullin, A., Gilmullin, R., Khakimov, B. *National Corpus of the Tatar Language "Tugan Tel": Grammatical Annotation and Implementation*. In: Procedia – Social and Behavioral Sciences. Vol.95 (2013). pp. 68-74.

Tatar National Corpus. http://corpus.antat.ru.

Derzhanski, I., Kotsyba. N. (2009). *Towards a Consistent Morphological Set for Slavic Languages: Extending MULTEXT-East for Polish, Ukranian and Belorusian.* In: Metalanguage and Encoding Scheme Design for Digital Lexicography. Proceedings of MONDILEX 2009. Ed.: R.Garabik. Bratislava, 2009. pp. 9-26.

Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A., Divjak, D. (2008). *Designing and Evaluating a Russian Tagset.* In Sixth International Conference on Language Resources and Evaluation, LREC'08, Paris, ELRA.

Resolution of the scientific-practical seminar "Unification of systems of grammatical annotations of Turkic languages corpora (seminar UniTurk)". URL: http://ips.antat.ru/page.php?id=225 (In Russian).

Tatar Grammar (1993). In 3 volumes. Kazan: Tatar publishing company, 1993. V. 2: Morphology. 397 p. (In Russian).

Tatar Grammar (2002). In 3 volumes. Moscow: Insan, Kazan: Fiker, 2002. V. 2. – 448 p. (In Tatar).

Plungyan, V. (2003). *General Morphology. Introduction.* Moscow: Editorial URSS, 2003. 384 p. (In Russian).

The Leipzig Glossing Rules. – URL: http://www.eva.mpg.de/lingua/resources/glossing-rules.php.

Dybo, A., Sheymovich, A. (2014). A*utomatic morphological analysis for corpora of Turkic languages*. In: Philology and culture. 2014. № 2. pp. 20-26. (In Russian)

Bashkir corpus. http://web-corpora.net/bashcorpus/search/index.php?interface_language=en.

Galieva, A., Khakimov, B., Gatiatullin, A. (2013). *A Metalanguage for Describing theStructure of Tatar Word Forms for Corpus Grammatical Annotations.* Uchenye Zapiski Kazanskogo Universiteta. Seriya Gumanitarnye Nauki, 2013, vol. 155, no. 5, pp. 287-296. (In Russian).

Kibrik, A., Arkhipov, A., Daniel, M., Kodzasov, S., Mayers, T., Nakhimovski, D. (2007). *Digital processing of linguistic data for minority languages documentation* URL: http://www.dialog-21.ru/digests/dialog2007/materials/html/35.htm (In Russian).

Atwell, E., Demetriou, G., Hughes, J., Schiffrin, A., Souter, C., & Wilcock, S. (2000). *A comparative evaluation of modern English corpus grammatical annotation schemes.* ICAME Journal: International Computer Archive of Modern and Medieval English Journal, 24, 7-23. pp. 7-23.

Leech, G. (2004). *Developing Linguistic Corpora: a Guide to Good Practice Adding Linguistic Annotation.* Oxford: Oxbrow Books. pp. 17-29.