TO THE PROBLEM OF UNIFICATION OF THE ANNOTATION SYSTEMS OF GRAMMATICAL CATEGORIES IN THE CORPORA OF TURKIC LANGUAGES

B. Khakimov
Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences
Kazan Federal University
khakeem@yandex.ru

A. Galieva
Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences
Kazan Federal University
amgalieva@gmail.com

A. Gatiatullin
Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences
Kazan Federal University
agat1972@mail.ru

ABSTRACT

This article presents the first results of the comparative analysis of grammatical annotation systems in different corpora of Turkic languages. The nominal and the verbal inflection is analyzed, as well as the meta-language of the formal description of grammatical categories.

1 Introduction

The development of Turkic studies during the past years has been marked by the deepening of the theoretical foundation of linguistic research, the increasing emphasis on new directions and challenges of modern linguistics, including applied linguistics. Criteria and principles of substantial analysis of inflectional categories remain one of the priority directions of research in Turkic languages, and the interest to this traditional subject is supported by formulation of new tasks, posed by the development of information technologies.

According to the developers' experience, representing the information about grammatical categories of Turkic languages in the corpus annotation is an independent scientific problem that intersects with different and sometimes opposite approaches to the description of grammatical phenomena.

One of the most important tasks of the development of grammatical annotation in the corpora of Turkic languages is to identify the inventory level of inflectional categories and to create an optimal meta-language for description of these grammatical categories.

2 General principles

It is known that for a long time the Turkic linguistics, especially in Russia, developed under the strong and direct influence of the Indo-European linguistics, when certain facts and even grammatical categories in the Turkic languages were often compared and linked not with each other, but with similar facts in the languages, which were theoretically more developed. Thus, distinguishing of a number of grammatical categories in the Tatar language took place under the direct influence of the Russian linguistics.

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 terms [1, p.23].

In the course of work on the system of grammatical annotation of the Tatar National Corpus (hereinafter, TatNC) [2; 3], we made an attempt to create a meta-language 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. There are special studies on the common meta-language and tagset, for example, for Slavic languages [4; 5]. In 2014, on the Uniturk workshop which was held in Kazan, this problem was discussed for the first time for Turkic languages [6].

To provide the resource with more flexibility, the developers of the TatNC, on the one hand, relied on the information provided in the academic grammars of the Tatar language [7; 8], and on the other – appealed to special studies on general morphology and linguistic typology ([9], and others).

Tags for parts of speech and grammatical categories were created to meet the worldwide standards, primarily the Leipzig glossing rules [10].

From the point of view of the user's convenience and the research problems to be solved, grammatical annotation in the corpus should meet the following requirements: 1) simplicity; 2) relevance to the grammatical and semantic system of the Tatar language; 3) transparency to the user – an average linguist-turkologist; 4) universality: it should be understandable for linguists who are not specialists in the field of the Tatar language, e.g. typologists; 5) compatibility with tagged corpora and lexicographic databases of other languages. Simultaneous satisfaction of all these requirements is a rather complicated task, since these precepts are often mutually contradictory.

Therefore, the designed system should provide the convenience of search, regardless of the users' theoretical assumptions. It means that regardless of the name or interpretation of a particular formation, the system must allow finding these forms in their real textual manifestation and investigating them.

Let us mention some of the factors that hinder the process of grammatical annotation of corpora for the Turkic languages in general:
1) poor differentiation of word-building and inflection in the Turkic languages, the lack of clear boundaries between them;
2) polysemy and homonymy of affixes;
3) lack of common standards on the reflection of linguistic information in electronic corpora and lexicographical databases;
4) lack of a single meta-language to refer to grammatical categories.

We have carried out a comparison with the grammatical annotation system of the corpora of Minority Turkic languages, which is being developed under the leadership of A. Dybo (hereinafter, MTLC) [11], as well as with the grammatical annotation system of the Bashkir language corpus (hereinafter, Bashmorph) [12].

3 Nominal Inflection

It has been observed that the above mentioned annotation systems have much in common.

Thus, the tags for the cases are the same.
Nom – the main case; it is not marked;
Gen – genitive;
Dir – directive;

Acc – accusative;
Loc – locative;
Abl – ablative.

In the MTLC system, there is distinction between Dat – dative and Dir – directive. In the modern Tatar language they correspond to a single case, and it is tagged in the TatNC as directive (Dir), which is a calque of the name that is traditionally used in the Tatar language grammars (yünäleş kileşe). The Bashmorph system employs the tag DAT.

In addition, MTLC system contains tags for a number of cases that are missing in the modern Tatar language: Equ – equative (comparative case, comparative-restrictive case, comparative-limiting case), Part – directive-partitive, Simil – similative, Comit – comitative.

TatNC annotation system has a special tag for the directive case with a limiting meaning (DIR_LIM), or limiting case. This form results from the fusion of the directive case affix with the affix –ça, and it occupies an intermediate position between a case form and an adverb (most turkologists refer such formations to adverbs). In the Tatar grammar book, this form is called “directive case with the affix -gaça/-gäçä” [7, p.51].

TatNC has special tags for attributive forms derived from nouns: ATTR_MUN – attributes ending on -lı (munitative); ATTR_ABES – attributes ending on -sız (abessive); ATTR_LOC – attributes ending on -dağı (locative); ATTR_GEN – attributes ending on –nıkı (genitive). It is noteworthy that in Bashmorph there is also a special tag ABE to indicate abessive.

The category of posessivity in the corpora is marked similarly (Table 1).
In MTLC the singular (SG) is not marked on its own; according to the developers, it is always cumulative with the person or possessiveness. In TatNC and Bashmorph there is a separate tag SG for singular, which is part of the integrated tags, related to posessiveness (see Table 1) or person (1SG – 1 person singular) and case (*kitapqa* 'to book' - (N) SG, DIR).

Table 1. Category of possessiveness

TatNC	MTLC	Bashmorph
POSS_1SG	Poss1 – 1 person posessor	POSS
POSS_2SG	Poss2 – 2 person posessor	POSS
POSS_3SG	Poss3 – 3 person posessor	POSS

In general, in the nominal annotation systems there are more common traits than differences.

4 Verbal Inflection

In the verbal annotation are also many similarities, but there is also a significant number of mismatches. Let us consider the representation of verb tenses in the compared annotation systems.

Table 2: Verb tenses in TatNC.

PRES	Present tense	-Y
PST_DEF	Past categorical tense	-DI
PST_INDF	Past resultative (perfect) tense	-GAn
FUT_DEF	Future categorical tense	-AçAk
FUT_INDF	Future indefinite tense	-[I]R
FUT_INDF_NEG	Negative form for future indefinite tense	-mAS

The tense system in Bashmorph is quite similar:

Table 3. Tense system in Bashmorph.

PRES	present tense
PST.DEF	past definite tense
PST.INDF	past indefinite tense
FUT.INDF	future indefinite tense
FUT.DEF	future definite tense

Verb tenses in MTLC have the following forms:
Sequ – sequentative, action prior to the main action;
Praes – present tense;
Fut – future tense;
Indir – indirective, past (perfect) tense with indirect evidentiality;
Perf – perfect tense (action in the past with result in the presence), unmarked by evidentiality;
Res – resultative tense (perfect tense, marked by direct evidentiality. It is either an action that was witnessed by the speaker, or an action, whose result is observed in the present);
Praet – preterite tense (past unmarked);
FutIm – immediate future tense.

In the verb tenses description in MTLC the emphasis is put on the evidentiality (when explaining the tags). In the course of development of the past tenses annotation for TatNC, one of the suggested options was also associated with the marking of evidentiality, but later we decided not to adopt it. To refer to the past and future tenses, we employ the tags that indicate the definiteness as the most simple and clear to a broad category of users.

Working on the annotation system for the Tatar corpus, we had great controversy regarding the annotaion of adverbial-participial forms. 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 the Tatar linguists, so it is unfamiliar to many people. Gerund (GER tag is used in Bashmorph) is familiar because of the foreign languages, but it does not correspond substantially to the respective category in the Tatar language, because it is used in the grammars of European languages. The expression "adverbial verb" is the translation of the Russian word *deeprichastiye* into English, and in its essence it is a kind of calque for the Tatar term *häl fiğıl*, but it is quite cumbersome for the corpus annotation. It is used in the current version of the TatNC 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 that the CONV tag have been used in one of the earlier annotation systems.

Great difficulties are caused by the most frequent forms of the adverbial participle ending on *-ıp*, because they embody a large complex of meanings, which are difficult to cover with a single "label". In the current version of TatNC annotation they are called "adverbial verb of a concurrent (accompanying) action".

In MTLC the adverbial participle is denoted as Conv (ConvFin – goal converb, ConvDelim – converb of limited action).

The action name in TatNC is referred to as VN – verbal noun, while in the Bashmorph as SUP – supine (we consider this term to be unsuitable because in many languages it indicates purpose). When the semantic difference between the forms of verbal nouns is unclear or difficult to formalize, TatNC uses digital indexes for tags: VN_1 – name of the action on -U; VN_2 – name of the action on -[I]ş.

In the compared annotation systems there are tags to denote the manner of verbal action, which is expressed by affixal means. TatNC distinguishes between the raritive form ending on -gala (RAR_1) and the raritive form ending on –ıştır (RAR_2). The Bashmorph system has the ACYCL tag to refer to an acyclic action.

Certain orthographic peculiarities are worth mentioning as well. Currently, we mark the present tense as PRES (orientation to new languages), while in MTLC it is marked with the tag Praes (with ae digraph, according to the Latin spelling).

5 Conclusion

This article provides the first results of the comparative analysis of annotation systems used in different corpora of Turkic languages developed in Russia. Further research is 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 standards and to create effective systems of automatic text processing for kindred languages.

6 References

[1] **GUZEV V. G., NASILOV D.M.** Inflectional categories in the Turkic Languages and the Concept of “Grammatical Category”]. In: Soviet Turkology, 1981. N 3. Pp. 22–35 (In Russian).

[2] National Corpus of the Tatar Language “Tugan Tel”: Grammatical Annotation and Implementation. In: Procedia – Social and Behavioral Sciences. Vol.95 (2013). Pp. 68-74.

[3] Tatar National Corpus. http://web-corpora.net/TatarCorpus/search/?interface_language=en.

[4] **DERZHANSKI I., KOTSYBA N.** 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.

[5] **SHAROFF S., KOPOTEV M., ERJAVEC T., FELDMAN A., DIVJAK D.** Designing and Evaluating a Russian Tagset. In Sixth International Conference on Language Resources and Evaluation, LREC’08, Paris, ELRA.

[6] 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). August 2014.

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

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

[9] **PLUNGYAN V.A.** General Morphology. Introduction. – Moscow: Editorial URSS, 2003. 384 p. (In Russian).

[10] The Leipzig Glossing Rules. – URL: http://www.eva.mpg.de/lingua/resources/glossing-rules.php. August 2014.

[11] **DYBO, A.V., .SHEYMOVICH A.V.** Automatic morphological analysis for corpora of Turkic languages In: Philology and culture. 2014. № 2. Pp. 20-26.

[12] Bashkir corpus. URL: http://web-corpora.net/bashcorpus/search/index.php?interface_language=en. August 2014.