***A.R* *Gatiatullin., D.S. Suleymanov,
N.A. Prokopyev, M.M. Saifullin***

**«Turkic Morpheme» portal as a tool for unification of annotation system for Turkic electronic corpora**

**Abstract.** In the present time, in connection with the urgent task of creating electronic corpora of natural languages for their preservation and development, there is a surge in the number of electronic corpora for the Turkic languages. Considering cooperation with corpora developers, as well as with native speaker experts, we propose a single conceptual space for Turkic corpora development through unification of the annotation system using the «Turkic morpheme» portal presented in this paper.

**Keywords.** Turkology, Corpus Annotation, Linguistic Resource, Frame Ontology.

**1. Introduction**

Relevance and extreme importance of creating electronic corpora of natural languages for their preservation and development, as well as for a resource base of NLP technologies and language research, is beyond doubt. In the present time, due to the growing interest, understanding and support from government, international institutions, foundations with great attention to the problems of endangered and low-resource languages from UNESCO, there is a surge in the number of electronic corpora for many languages, including Turkic languages [UNESCO LT4All 2019, TurkLang 2020]. For many of languages from the Turkic family several electronic corpora are already developed. The following is a non-exhaustive list of Turkic languages corpora of those resources that we refer to in our research: Turkish National Corpus (TNC) – [www.tnc.org.tr](http://www.tnc.org.tr); Almaty Corpus of Kazakh language (NCKL) – [web-corpora.net/KazakhCorpus](http://web-corpora.net/KazakhCorpus/search/); Corpus of the Altai language – [altay2.gasu.ru](http://altay2.gasu.ru/); National Corpus of the Bashkir language – [bashcorpus.ru](http://bashcorpus.ru); Bashkir Poetic Corpus – [web-corpora.net/bashcorpus](http://web-corpora.net/bashcorpus/search/); «Tugan Tel» Tatar National Corpus – [tugantel.tatar](http://tugantel.tatar); Corpus of Written Tatar – [www.corpus.tatar](http://www.corpus.tatar); Corpus of the Khakass language – [khakas.altaica.ru](http://khakas.altaica.ru); Corpus of the Yakut language – [adictsakha.nsu.ru/corpora/corp](http://adictsakha.nsu.ru/corpora/corp); Сorpus of the Uzbek language – [corpus-uz.herokuapp.com](http://corpus-uz.herokuapp.com); Digital Corpora in Siberian Minority Languages (Teleut and Shor) – [corpora.iea.ras.ru/corpora](https://corpora.iea.ras.ru/corpora/).

In cooperation with the corpora developers [Abduraxmonova, 2021], as well as with native speaker experts, we propose a single conceptual space for Turkic corpora development through unification of the annotation system using the «Turkic morpheme» portal presented in this paper.

**2. Analysis of Turkic languages electronic corpora**

The development of electronic linguistic corpora presents developers with a wide range of problems, successful solution of which requires combining the results of linguistic research and modern computer methods of linguistic data analysis. Capabilities of a corpus are largely determined by annotation system used in it.

In the context of globalization and integration of scientific research, the issues of linguistic data representation unification in corpora acquire special significance – systems for grammatical categories annotation in particular; this is especially important for groups of related languages. Analysis of the current situation in Turkic corpora (according to articles [Aksan et al. 2018, Kubedinova et al. 2019]) shows that in Turkic linguistics, despite genetic and structural-typological commonality of the Turkic languages, general principles and approaches to linguistic annotation of texts have not yet been formed. Obviously, this will lead to significant difficulties in comparative studies in the future, as well as in development of Turkic parallel corpora, multilingual text processing systems, and in solving of other theoretical or applied problems.

Differences in annotation systems concern both grammatical categories inventory with meta-language of their description, and composition of the required data representation layers. The same morphological categories in different studies on the Turkic languages are annotated differently. Corpus developers often use annotation systems created for Indo-European languages, which do not always adequately and fully reflect the specific features of the Turkic languages, therefore, the development of an annotation system for this particular language group is a very relevant problem.

The lack of uniformity in corpora annotation is associated with objective scientific and organizational problems. The organizational form for the creation of a unified annotation system is to hold joint conferences and seminars, to discuss the unification issues for the Turkic languages, and to implement joint projects. One of such events is the workshop on unification of grammatical annotation systems in the Turkic languages corpora (UniTurk), which is regularly held within the framework of international conferences TEL and TurkLang. At UniTurk (Kazan 2014, Istanbul 2014, Kazan 2015, Bishkek 2016), the problems of developing a unified morphological annotation of Turkic languages texts for use in corpora and automatic text processing systems were discussed.

Despite the organizational measures, practice has shown that these alone are not enough, the problem of unification is being solved rather slowly and ineffectively. Unified linguistic resources and tools are needed in addition with online platforms for discussing these issues. «Turkic Morpheme» portal ([modmorph.turklang.net](http://modmorph.turklang.net/en/), [Gatiatullin et al. 2020]) was proposed as linguistic resource with a set of services and as a platform for communication on the problem.

**3. Description of the «Turkic Morpheme» portal**

The first stage is the unification of grammatical categories expressed with morphemes in Turkic languages, both affixal and root morphemes. A glossary of grammemes which express grammatical values has been developed for this purpose (Fig. 1). This glossary is available to all processing annotation data processing software for Turkic languages, in particular, to morphological analyzers that will be used for Turkic electronic corpora unified annotation. A detailed description is given for each grammeme in several languages (Fig. 2). This description is necessary for language experts and developers to unambiguously perceive the annotation system.



*Fig. 1.* **Fragment of grammeme glossary on portal**

The portal has a set of pivot tables that provide functions of database overview for all languages. In particular, these tables make it possible to assess the degree of grammatical values expression in linguistic units in different Turkic languages and to compare Turkic languages according to the degree of grammatical affinity.

The assignment of grammatical values to linguistic units (affixal and root morphemes, postpositions and postpositions) is actually based on the language expert's intuition. One linguistic unit can be assigned to several of the most frequent grammatical values. For example, the morpheme -GA in the Tatar language expresses two values: dative and directive. In the Yakut language, these grammemes are expressed by two different affixal morphemes.



*Fig. 2.* **Description of the grammeme on portal**

In order to see a more complete picture of grammatical values representation, statistical studies on corpus data are needed. However, morphological analyzers and statistical analysis software do not have the appropriate functions for obtaining such data; they only allow to part out the linguistic units from word-forms. To implement these functions, it is necessary to apply analyzers of semantic-syntactic level.

It is not possible to use the currently popular machine learning technologies for these problems, since most of the Turkic languages do not yet have sufficient linguistic resources to teach them. Therefore, it is important to create the appropriate necessary resources. Moreover, when creating such resources, it is also necessary to observe the principle of annotation system universality within the group of Turkic languages. Obviously, the principle of universality for a language group is easier to observe, having uniform multilingual resources and services for these languages. The proposed «Turkic morpheme» portal can be used as such multilingual resource and set of services.

**4. Frame ontologies**

Frame ontologies have recently been quite actively used in various subject areas for knowledge processing problems. They serve, on the one hand, as a knowledge base for semantic-syntactic analyzers that are used to annotate the electronic corpora [Yelibayeva 2020], and on the other hand, they are used for language comparison. For example, the Turkic languages have different sets of affixal morphemes, which partially coincide, and the percentage of this coincidence allows expert to determine the grammatical proximity of languages and to classify them into subgroups: Kypchak subgroup, Oguz subgroup, Karluk subgroup, etc.

Some of the most complete and well-known electronic resources that provide semantic role models are FrameNet, VerbNet [Kipper et al. 2006] and PropBank [Palmer et al. 2005]. A similar resource developed for the Russian language is the FrameBank [Lyashevskaya et al. 2015]. It combines a dictionary of lexical constructs with annotated corpus of their expression in texts of the National Corpus of Russian language. FrameBank constructs include predicate-argument structures for verbs, nouns, adjectives, adverbs, and predicatives. The only similar resource for the Turkic languages known to authors is Turkish PropBank [Sahin 2016].

To define a frame ontology, the model proposed in [Avdeenko et al. 2013] is used:

OF = <C, R, S, G, T, DS, DG, E>.

C – set of frame classes describing concepts of ontology domain.

R – a set of binary relations on frame classes, R = {RISA} Ս {RASS}, where RISA – is a set of «class-subclass» hierarchical relations; RASS – is a set of associative relations.

S – a set of slots (class attributes).

G – a set of facets (slot attributes).

T – a set for controlled dictionary of ontology domain terms.

DS – a set of slot types.

DG – a set of facet types.

E – a set of class individuals.

The peculiarity of frame ontology, implemented in «Turkic Morpheme» portal, is that it is initially multilingual, therefore the model is built similarly to FrameNet around situations. The situation is represented in form of a frame class, which is common to all languages. The structure of the portal database part that implements this frame ontology is shown in Fig. 3 and 4.

Thus, the complete database of situational frames has a number of applications for Turkic languages corpora development. These include:

1. System of standards for semantic-syntactic annotation;
2. Linguistic resource for semantic-syntactic analyzers development, including the subsequent use in machine learning.



*Fig. 3.* **Common part of the frame ontology database**



*Fig. 4.* **Language-specific part of the frame ontology database**

**5. Conclusion**

The multilingual linguistic database, presented in the «Turkic morpheme» portal, should become an electronic resource that will allow unifying of linguistic resources for Turkic languages. The presented language constructs are to be linked with examples in the Turkic languages electronic corpora. Also, a set of linguistic services will be developed on the basis of the portal according to unified NLP pipeline standard.

**References**

1. *Aksan M., Aksan Y.* (2018), Linguistic corpora: A view from Turkish, Studies in Turkish Language Processing. Turkish Natural Language Processing. Theory and Applications of Natural Language Processing. Springer, Cham, pp. 301–327.

2. *Avdeenko T.V., Bakaev M.A.* (2013), Gibridnaya model' predstavleniya znaniy dlya realizatsii vyvoda vo freymovoy ontologii [Hybrid model of knowledge representation for inference in frame ontology], Nauchnyy vestnik NGTU [Scientific Bulletin of NSTU], no. 3, pp. 84–90.

3. *Abduraxmonova N.* (2021), O‘zbek tili korpusini yaratishda lingvistik annotatsiyalash tamoyillari, International Journal of Word Art, vol. 4, is. 1, pp. 164-174.

4. *European Language Resources Association* (2019), Collection of Research Papers of the 1st International Conference on Language Technologies for All (UNESCO LT4All 2019).

5. *Gatiatullin A., Suleymanov D., Prokopyev N., Khakimov B.* (2020), About turkic morpheme portal, Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020). CEUR-WS.org, pp. 226–243.

6. *Institute for history, language and literature, Ufa scientific center of Russian Academy of Sciences* (2020), Proceedings of the VIII International Conference on Computer Processing of Turkic Languages (TurkLang 2020).

7. *Kubedinova L., Adali E.* (2019), The Crimean Tatar Electronic Corpus vs the Electronic Corpus of the Turkish Language (Grammatical Tagging of Noun, Verb), Proceedings of 4th International Conference on Computer Science and Engineering (UBMK-2019), pp. 783–788.

8. *Kipper K., Korhonen A., Ryant N., Palmer M.* (2006), Extending VerbNet with novel verb classes, Fifth International Conference on Language Resources and Evaluation (LREC 2006), pp. 1027–1032.

9. *Lyashevskaya O., Kashkin E.* (2015), FrameBank: A Database of Russian Lexical Constructions, Analysis of Images, Social Networks and Texts. AIST 2015. Communications in Computer and Information Science, vol 542. Springer, Cham, pp. 350–360.

10. *Palmer M., Gildea D., Kingsbury P.* (2004), The Proposition Bank: A corpus annotated with semantic roles, Computational Linguistics Journal, 31(1), pp. 71–105.

11. *Sahin, G.G.* (2016), Framing of verbs for Turkish PropBank. Proceedings of 17th International Conference on Intelligent Text Processing and Computational Linguistics, (CICLING 2016).

12. *Yelibayeva G., Sharipbay A., Mukanova A., Razakhova B*. (2020), Applied ontology for the automatic classification of simple sentences of the Kazakh language, Proceedings of 5th International Conference on Computer Science and Engineering (UBMK-2020), pp.13-18.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Gatiatullin Ayrat**

Institute of Applied Semiotics of TAS (Russia).

***E-mail:*** ***ayrat.gatiatullin@gmail.com***

**Suleymanov Dzhavdet**

Institute of Applied Semiotics of TAS (Russia).

***E-mail:*** ***dvdt.slt@gmail.com***

**Prokopyev Nikolai**

Institute of Applied Semiotics of TAS (Russia).

***E-mail:*** ***nikolai.prokopyev@gmail.com***

**Saifullin Miras**

Kazan Federal University (Russia).