About Turkic Morpheme Portal

Ayrat Gatiatullin[1,2][0000-0003-3063-8147], Dzhavdet Suleymanov[1,2][0000-0003-1404-0372], Nikolai Prokopyev[1,2][0000-0003-0066-7465], and Bulat Khakimov[1,2][0000-0003-0126-9522]

[1] Institute of Applied Semiotics of Tatarstan Academy of Sciences, Kazan, Russia
[2] Kazan Federal University, Kazan, Russia
nikolai.prokopyev@gmail.com

Abstract. Doing scientific research in fields of turkology and agglutinative languages typology requires software that takes into account the structural and functional features of languages in question. This paper presents a description of the Turkic Morpheme Portal, in the development of which an integrated approach is used for the development of computer linguistic models and technologies for Turkic languages processing. This portal was created on the basis of the structural-parametric functional model of the Turkic morpheme and contains special linguistic databases that describe the categories of Turkic languages at different levels: morphological, syntactic, and semantic. The problems of developing complex multilingual linguistic models for low-resource languages and their software implementation are considered. The prospects of using the created portal as a base for the development of linguistic software, as well as an information and reference system, including a multilingual thesaurus, and as a platform for communication of specialists are given.

Keywords: Linguistic resource, Ontology, Thesaurus, Turkology, Multilingual model, Morphology.

1 Introduction

The importance of Turkic Morpheme Portal development is determined by the situation that has established around development of software for Turkic languages processing, which is formed with a number of factors.

Firstly, historically, development of linguistic software for English is leading in field of computational linguistics in the world, and in Russia, in turn, it is software for Russian, therefore, researchers and developers of scientific software for other languages study technologies for English and Russian, basing their research on these methods. However, the Turkic languages, unlike Russian and English, fully belong to agglutinative languages family and are structurally quite different. This means that computational linguistic models specialized on the agglutinative languages and technologies for processing of this type of languages are needed.

Secondly, all Turkic languages, except for Turkish, are low-resource languages and their lag behind the resource-rich languages continues to accumulate. One of the rea-

sons for this is lack of specialists working on development of linguistic resources and computer processing software for Turkic languages.

Thirdly, despite the increase in research on Turkic languages processing, there is practically no real integration of the results. According to resolution of TurkLang 2017 Conference, joining efforts of groups working on different Turkic languages can provide solution to this problem.

Fourthly, there is a duplication of linguistic models and resources, as well as language processing software, which are basically 70-80 percent common to all Turkic languages. Therefore, overcoming such duplication and combining the efforts in collaborative development and interchange of linguistic software is urgent.

Fifthly, the idea of creating a machine fund of Turkic languages was proposed back in 1988 in the "Soviet Turkology" journal by V.G. Guzev, R.G. Piotrovsky, A.M. Shcherbak in their article "On the Creation of the Machine Fund of Turkic Languages" [1]. In this article, a number of ideas about the principles of machine fund organization were presented, which remain relevant at the present time. However, for a number of reasons, these ideas were never implemented. We'll look at them in detail in the next chapter.

2 Related work

2.1 Structure of NLP domain

Current state of NLP research domain is characterized by rapid development of machine learning and neural networks technologies. A problem with machine learning systems is the need for large amounts of data with different types of annotation, such as morphological, semantic, and syntactic. The development of machine learning methods has led to the fact that a number of problems previously solved using ontologies, frames and semantic networks, began to be solved without these resources. One of these methods consists in constructing of vector representations for words in a low-dimensional space, namely word2vec [2], which allows to reduce the problem of words semantic proximity evaluation to calculating the cosine of the angle between the corresponding word vectors. Today this method is widely used for automatic construction and expansion of semantic resources, as well as in classification and clustering tasks.

Despite the advances in machine learning, high quality ontology models, frames, and semantic networks are still important linguistic resources. They are indispensable when high accuracy is required, even if it is achieved by narrowing the lexical coverage. One of the tasks in which linguistic resources built by experts like WordNet [3], are still out of competition is word sense disambiguation. Also, the ontological technologies and semantic networks are effective in evaluation of methods and systems for natural language processing. For these tasks such resources are used as a gold standard against which a comparison in some metrics is made.

Machine learning methods and ontological models represent two main directions for artificial intelligence development. Here, neural networks imitate empirical thinking and perception, while ontological models express logical and abstract thinking.

Machine learning methods require large corpora, and recently a number of corpora for Turkic languages were developed including but not limited to [4-6]. Other examples can be found at [7]. However, most of them cannot go beyond morphological markup, precisely because of the lack of ontological resources and software for syntactic and semantic analysis.

2.2 Linguistic resources

Multilingual databases and software constitute an effective typology tool for different languages, language units, properties, and phenomena classification. An example of such tool is “Global Lexicostatistical Database” [8]. This database presents the basic vocabulary of the world's languages in the comparative manner. It is intended for the formation of unified system of basic vocabulary lists to research the degree of world's languages relations based on the percentage of common words. Then the software system makes a classification genealogical tree of world languages.

Another function set that increases the efficiency and presentability of research software can be the combination of a linguistic service functions with geographic information systems. As an example of such service, one can propose a new resource for Turkic languages called “Maps for Turkic languages” [9], which is being developed by a group of Russian turkologists.

There were various attempts to combine different types of linguistic resources into a common database. One of these projects is BabelNet [10], a unified linguistic resource that combines 47 resources, including Wikipedia, WordNet, Wikidata, FrameNet, VerbNet, ImageNet and others. The integration of these resources in BabelNet is done automatically using a linking and lexical gap filling algorithm. It contains Babel synsets, presented in many languages and connected by a huge number of semantic relations: in version 4.0, out of 832 million meanings, more than 6 million concepts and 9.5 million named entities, linked by more than 1 billion semantic relationships are extracted.

Another example of linguistic resources unification is the SemLink project [11], which was developed at the University of Colorado. The authors of this project propose an approach to unification of the following resources: PropBank, VerbNet. FrameNet, WordNet.

3 Methodology

3.1 Problem definition

The analysis shows that among all the Turkic languages in all global international projects for linguistic resources, only one actively participated which is Turkish language. As a result, almost all Turkic languages, except Turkish, belong to low-resource languages. Therefore, an urgent task is formulated: to combine various kinds of linguistic resources for Turkic languages in one resource. When solving this problem, a hypothesis is put forward, that linguistic resources for the Turkic languages can be combined using the Turkic morpheme as a unifying element.

The choice of this element is based on structural features of Turkic languages. V.A. Plungyan [12] divides all languages according to morphological models into three types:

1. Elemental-combinatorial (Item and Arrangement) morphological model. The main structural tool of this model is linear segmentation.
2. Elemental-procedural (Item and Process) morphological model. In languages with this model some allomorphs are considered as initial, and others as derivatives, which can be obtained from the former by applying operations of "phonological processes".
3. Verbal-paradigmatic (Word and Paradigm) morphological model. In this model, there is a complete rejection of morphemic division when describing inflection. It is the word form that is chosen to be the minimum unit of grammatical description here.

According to this classification, the Turkic languages belong to elemental-combinatorial type. This allows us to consider the Turkic morphemes as integral elements in the Turkic language system, which are in different types of relationships with each other and with other elements of the language and semantics.

The next argument to the choice of Turkic morpheme as a basic connecting element is the following fact. The phonological and morphological levels of language contain a finite number of units, and it is possible to compose the finite alphabet of language in these units. At the same time, the syntactic and semantic levels of language operate with an infinite number of language units, which complicates the task of constructing a linguistic model.

In this regard, it is possible to extract the meanings of reproducible language units, i.e., units stored in memory in a complete form. For this, only two types of values can be distinguished:

- Meanings of morphemes;
- Meanings of words (including meanings of phraseological units).

B.Y. Gorodetsky proposes the folowing levels of analysis of two-sided language units [13]:

- Morpho-semantic level, represented by meanings of all morphemes distinguished in a given language;
- Lexico-semantic level, represented by meanings of all lexical units included in the lexicon of a given language.

Units on each level connect using morpho-semantic and lexical-semantic relations. The feature of Turkic languages is that a lexeme can coincide with a root morpheme. As a result, in Turkic languages both the unit of morpho-semantic level and the unit of lexico-semantic level are essentially the same, it is a morpheme. This determines the choice of the Turkic morpheme as the basic unit used to connect linguistic models of different levels.

3.2 Turkic Morpheme Model

On the basis of the stated hypotheses, a complex linguistic Turkic Morpheme Model was developed (Fig. 1). This model is a further development and generalization of the structural and functional model of Tatar affixal morpheme [14]. The original version was expanded onto all Turkic languages with inclusion of root morphemes. It is assumed that the resulting complex linguistic model can improve the efficiency of multilingual word processing software development. It also serves as a basis for solving other fundamental and applied problems, which require conceptual and formal linguistic models, common databases, as well as software based on these models. This is facilitated by the pragmatically oriented approach proposed by D.S. Suleymanov [15].

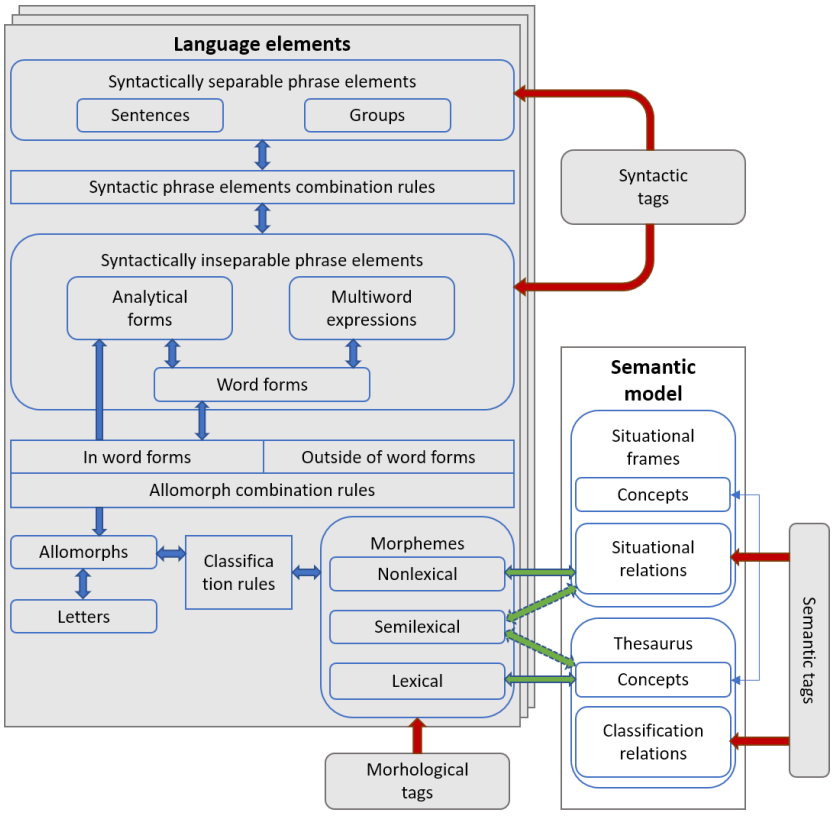


Fig. 1. Turkic Morpheme Model

3.3 Turkic Morpheme Portal

On the basis of the proposed model, the Turkic Morpheme Portal was developed. Web portal is a web site that provides access to various services in a particular area. The same way, the Turkic Morpheme Portal is a set of services for the Turkic languages processing. The portal has a whole range of functions:

1. An information and reference system on the Turkic languages grammar;
2. A place for communication of specialists in computer processing of Turkic languages for joint turkology research;
3. A set of linguistic resources for Turkic languages, including multilingual thesaurus and frame ontologies;
4. A database for linguistic services, such as text processors on different levels of language structure, presented as a pipeline.
5. A service for unification and connection of linguistic resources for Turkic languages, including corpora.
6. A source of linguistic data on Turkic languages for training the machine learning models.

The main purpose of the portal is supporting the research and development in the field of Turkic studies. This feature determines the requirements for the portal as a multipurpose scientific resource. The main properties of a scientific resource are as follows:

1. Multifunctionality. The resource is applicable to various tasks in which Turkic languages texts processing is required (machine translation, multilingual search, question-answer systems, information extraction).
2. Multilingualism. Resource software and algorithms are separated from linguistic database, while being focused on processing of Turkic languages, and are equally applicable to any language in this group.
3. Pragmatic orientation. The software is precisely oriented towards the processing of languages from Turkic family, it is not universal.
4. Stratification. The sentence analysis algorithms are based on representations of the sentence at several linguistic levels from morphological to syntactic and semantic.
5. Interactivity. Human-machine dialog interaction is required to resolve complex cases of ambiguity.

These principles were taken as a basis for Turkic Morpheme Portal development.

4 Portal description

4.1 Portal infrastructure

The Turkic Morpheme Portal has wide functionality available in two modes of portal operation: the reader mode and the expert mode, which actually represent two subsystems of the portal. The reader mode represents the reference subsystem, and the expert mode represents the research subsystem.

The reader can view the materials from system database, but they do not have the permission to edit it (Fig. 2). The portal provides the reader with descriptions of Turkic language elements collected by turkology specialists at different language levels, the rules for their combination (morphotactics), as well as the expressed meanings (semantic). The feature of this linguistic information presentation is that it is concen-

trated, classified and formalized in a single database, as well as equipped with options for searching and filtering queries.

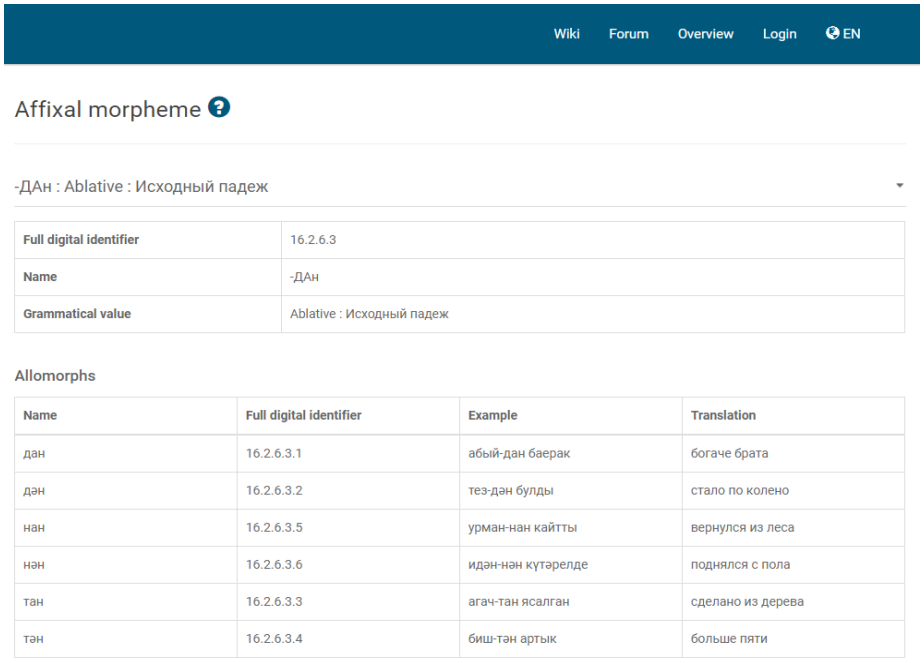


Fig. 2. Affixal morpheme page in the reader mode

The expert mode (Fig. 3) provides the user with a wide range of functions for carrying out research activities such as: collecting core information, classification, comparative statistical analysis, hypotheses verification. To take advantage of the expert mode, it is necessary for the user to pass authorization and get confirmation from the administrator, after which they get access to work with one selected language of their choice. The expert can also view the other languages in the reader mode. Within the framework of his chosen language, the expert has the permission to fill in the language-specific part of the model as well as some elements of the common part.

The third mode of portal usage, which is inaccessible for ordinary users, is the administrator mode (Fig. 4). It gives direct access to the database for any of the Turkic languages presented in the portal, as well as to the infrastructure part. The administrator confirms the expert roles, gives them the permission for portal translation, has access to the system log and is engaged in system technical support. Administrator mode is implemented using the standard framework tools.

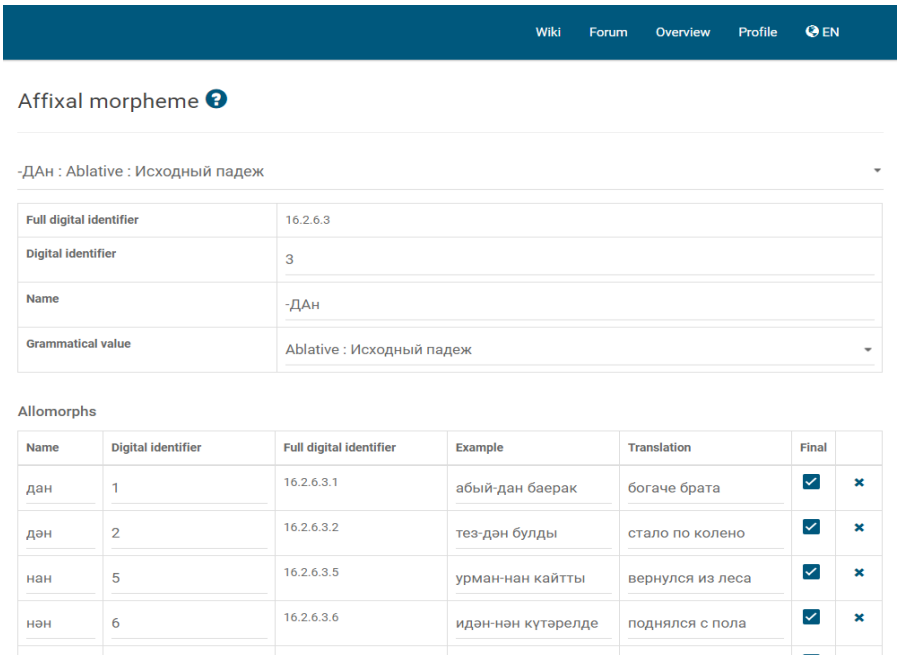


Fig. 3. Affixal morpheme page in expert mode

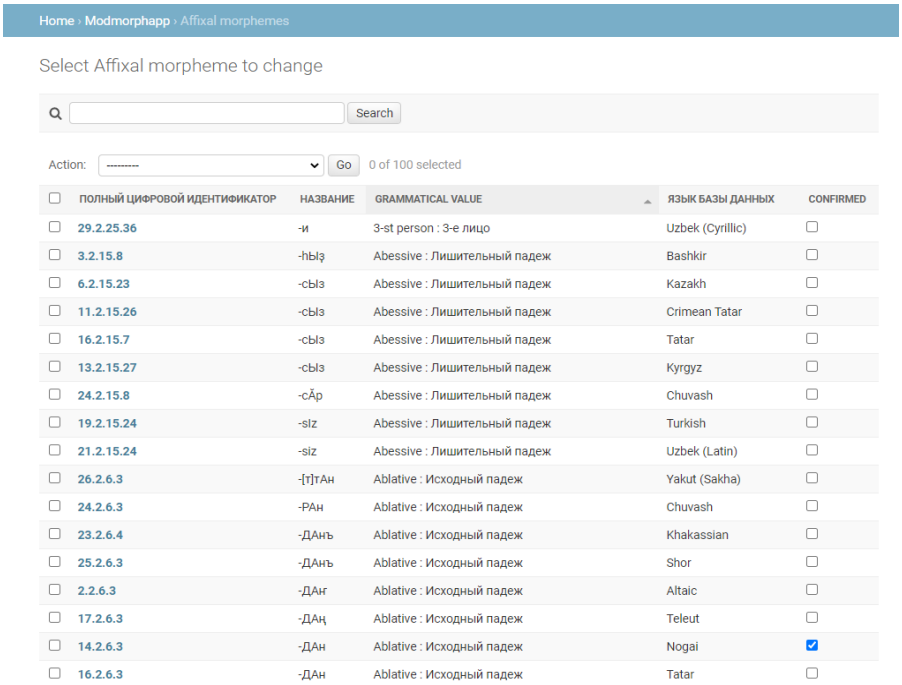


Fig. 4. Admin panel, affixal morphemes page

As a service with multilingual Turkic Morpheme Model support, the portal must have a multilingual interface. The portal framework toolbox provides a localization mechanism which allows users to translate the portal into other languages using the localization module (Fig. 5).

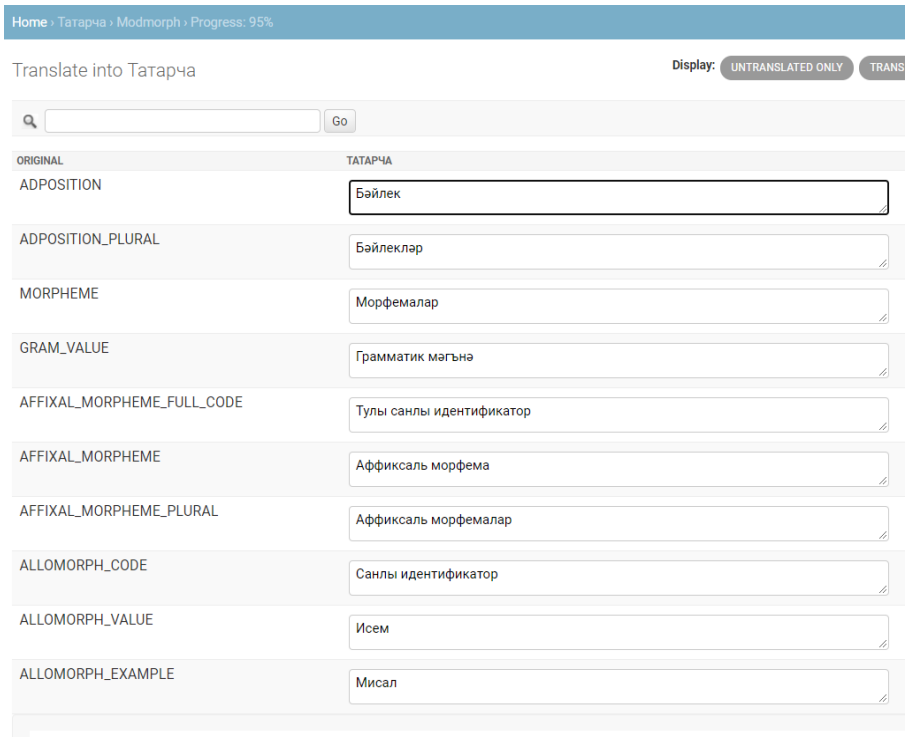


Fig. 5. Localization module, Tatar language example

The portal provides a forum for collaborative discussions and for user feedback. It also implements a wiki-like system that gives the user additional information about the model and the linguistic terminology. In addition, tools for summary overview are implemented, such as statistics of database and summary tables (Fig. 6), which provide an interlanguage representation of current model state with HTML and Excel formats.

Grammatical values + Morphemes (Download)

	Altaic	Azerbaijani	Bashkir	Chalkan	Chulym	Chuvash	Crimean Tatar	Dolgan	Gagauz	Khalaj	Karachay-Balkar
Abessive : Лишительный падеж			-һыҙ			-сĂр	-сыз				
Ablative : Исходный падеж	-ДАн'	-dAn	-ДАн			-РАн	-ДАн		-DAn		-дАн
Accusative : Винительный падеж	-Н[Ы]	-[n]I	-НЫ				-ны	-ны	-[y]I		-ны
Approximative : Приблизительное числительное			-лАп -лАҒАн								
Causative : Понудительный залог			-т -ДЫр			-[т]ТАр	-т -ДЫр				
Collective : Собирательное числительное			-АУ								
Comparative : Сравнительная степень			-[Ы]рАК			-[тА]рАх	-[д]ЖА				
Conditional Mood : Условное наклонение			-hA				-cA				
Conjunction : Союз											

Fig. 6. Summary table, grammatical value on morpheme example

4.2 Common part

The common part of the portal database contains the data for linguistic categories that are common to all languages presented in the model, such as: grammatical categories, grammatical values (grammemes, quasigrammemes, derivatemes) and concepts (objects, actions, attributes). Fig. 7 shows a conceptual database diagram for the common part of the model. Concepts here express some meanings and are arranged in a thesaurus with different relations between them. Grammatical categories and values express some linguistic modifiers, while grammatical values can be composite, i.e., consist of several nested values.

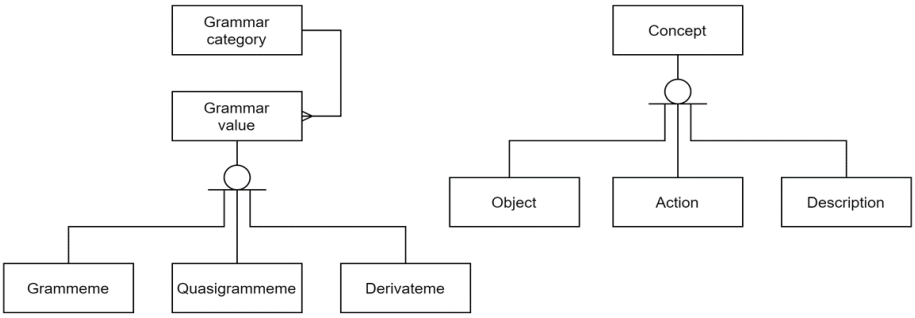
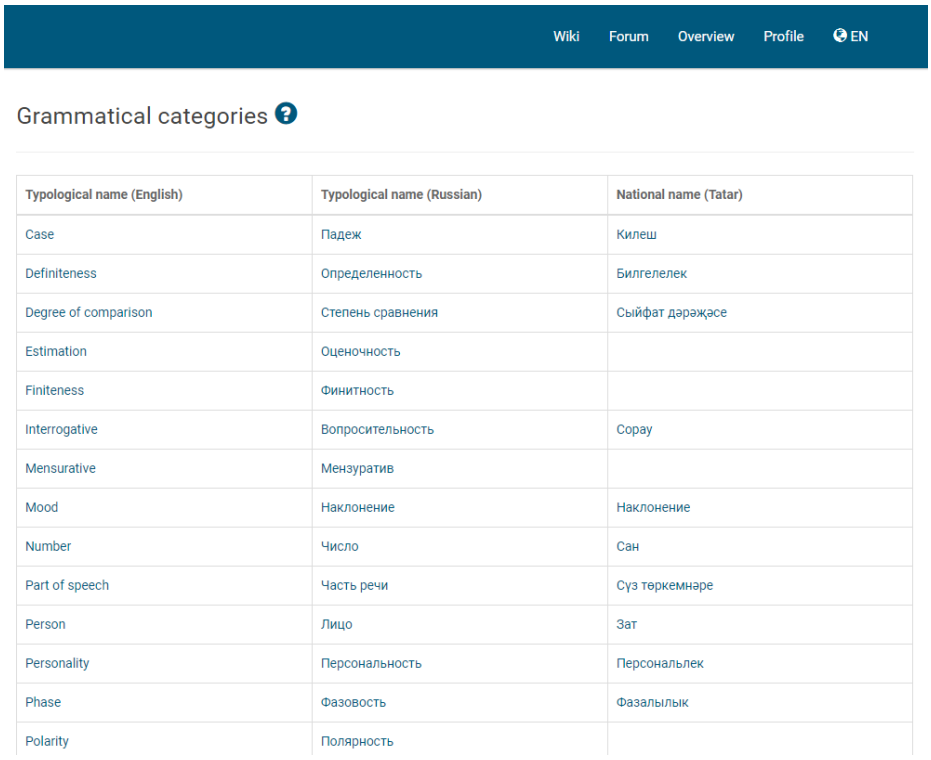


Fig. 7. Conceptual ER-diagram of the common part

As an example of working with the common part, let's consider the procedure of filling in the grammatical categories. By clicking on the left menu item “Grammatical categories”, the user gets access to the list of grammatical categories (Fig. 8).



Typological name (English)	Typological name (Russian)	National name (Tatar)
Case	Падеж	Килеш
Definiteness	Определенность	Билгелелек
Degree of comparison	Степень сравнения	Сыйфат дәрәҗәсе
Estimation	Оценочность	
Finiteness	Финитность	
Interrogative	Вопросительность	Сорау
Mensurative	Мензуратив	
Mood	Наклонение	Наклонение
Number	Число	Сан
Part of speech	Часть речи	Сүз төркемнәре
Person	Лицо	Зат
Personality	Персональность	Персональлек
Phase	Фазовость	Фазалылык
Polarity	Полярность	

Fig. 8. Grammatical categories list

The attributes “Typological name (English)” and “Typological name (Russian)” are filled in by the administrator, and users, both readers and experts, can only view them. The “National Name” field for a specific language, is intended to be filled by an expert. To enter information about a category in the expert’s language, he needs to select the category of interest in the list, after which a view of this category will open, where a form for national name editing is available (Fig. 9). In addition to the national name itself, an expert can enter the category description in their language and the source from which such a description is taken. In reader mode these items are provided read-only.

Work with other linguistic elements of the common part is done in a similar way, that is, their main part is filled in by portal administration, and then the expert has an opportunity to clarify the linguistic nuances for the corresponding Turkic language. In addition, a separate user role of typologist expert is set. These users can add their own concepts to the database (Fig. 10).

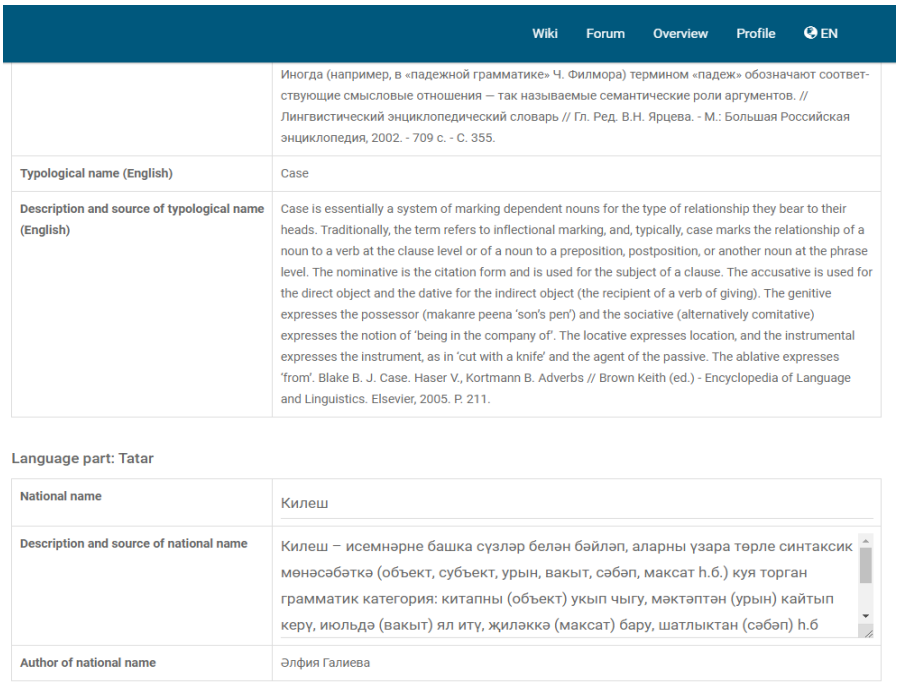


Fig. 9. Grammatical category view in the expert mode

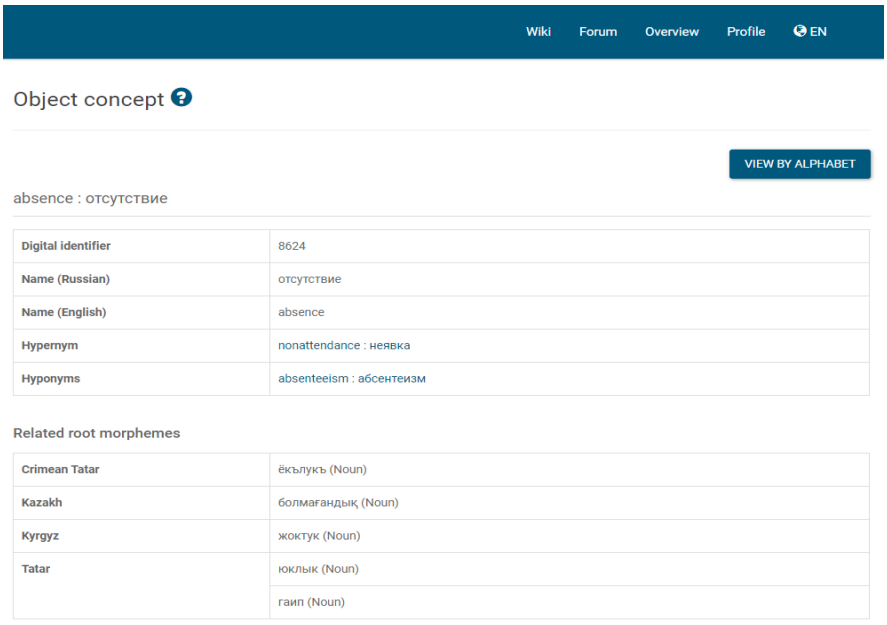


Fig. 10. Object concept view

4.3 Language-specific part

The language-specific part of the model includes those linguistic categories that completely depend on the specifics of a particular language. The following categories are implemented: affixal morphemes, analytical morphemes (particles, adpositions, auxiliary verbs), root morphemes and morphotactic rules. Affixal and analytical morphemes express some grammatical values, and root morphemes correspond to concepts from the common part. Morphotactic rules between root and affixal morphemes specify possible sequences of roots and allomorphs connections, and those between two affixal morphemes specify possible connections between pairs of corresponding allomorphs. Fig. 11 shows the conceptual database diagram for the language-specific part of the model.

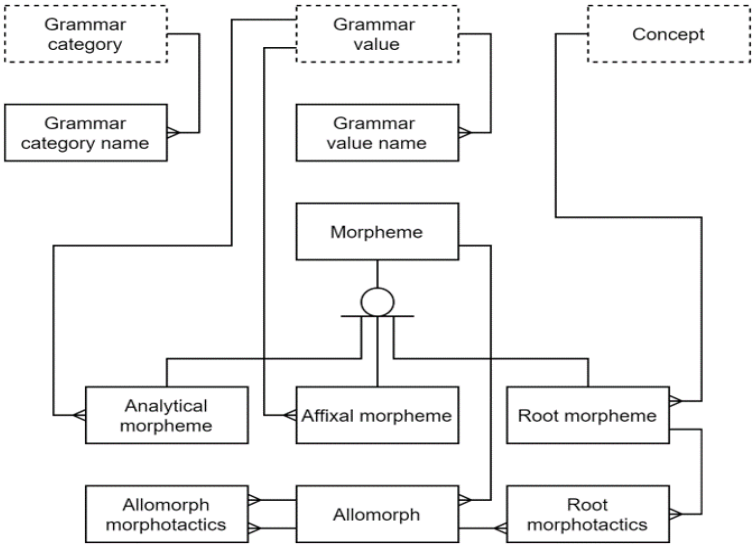
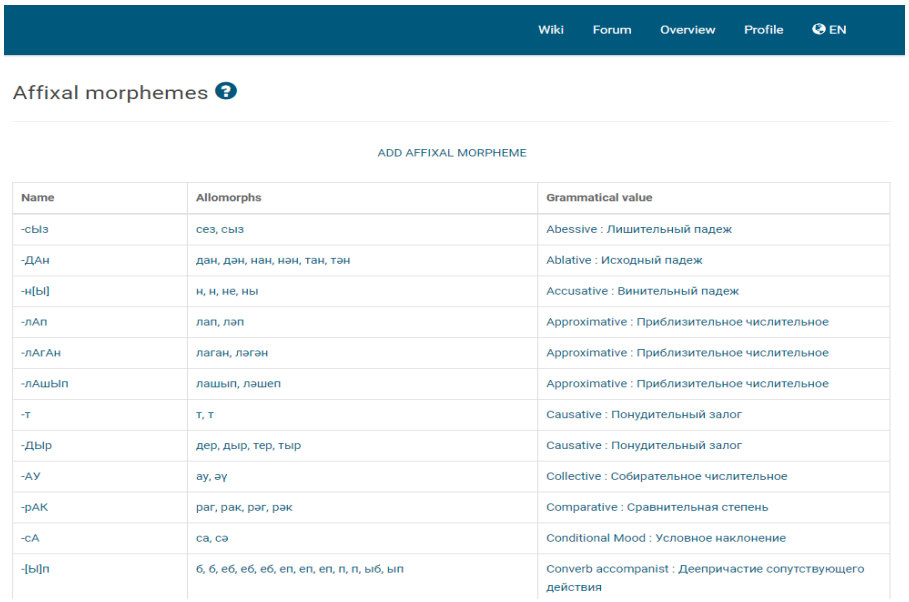


Fig. 11. Conceptual ER-diagram of the language-specific part

The expert has full access to the language-specific part for their language, they can add, edit, or delete elements of this part. As an example, filling the category "Affixal morpheme" is presented. Having selected the corresponding element from the left menu, the expert goes to a page with affixal morphemes list from system database (Fig. 12), where they can also add a new affixal morpheme.



Name	Allomorphs	Grammatical value
-сыз	сез, сыз	Abessive : Лишительный падеж
-ДАн	дән, дән, нан, нән, тан, тән	Ablative : Исходный падеж
-н[ы]	н, н, не, ны	Accusative : Винительный падеж
-лАп	лап, ләп	Approximative : Приблизительное числительное
-лАгАн	лаган, ләгән	Approximative : Приблизительное числительное
-лАшЫп	лашып, ләшеп	Approximative : Приблизительное числительное
-т	т, т	Causative : Понудительный залог
-ДЫр	дер, дыр, тер, тыр	Causative : Понудительный залог
-АУ	ау, әү	Collective : Собирательное числительное
-рАК	раг, рак, рәг, рәк	Comparative : Сравнительная степень
-сА	са, сә	Conditional Mood : Условное наклонение
-[ы]п	б, б, еб, еб, еб, еп, еп, п, п, ыб, ып	Converb accompanist : Деепричастие сопутствующего действия

Fig. 12. Affixal morphemes list

When adding or editing an existing affixal morpheme, the following attributes are available for filling in the morpheme data:

1. Textual value of the morpheme;
2. Grammatical value to which the morpheme corresponds;
3. Digital identifier for use in linguistic software.

Further, the expert can add specific variants of affixes for the morpheme called allomorphs. Each allomorph contains such attributes as: value, digital identifier, usage example, example translation, finality flag for morphotactics. The user interface for allomorphs is implemented in a table form. Figure 3 presented previously shows the entire form for editing the affixal morpheme. Editing of root morphemes is implemented in the same way.

Morphotactic rules, however, have a different editing interface. For the "Root+Affix" morphotactics, morphonological type database entity is used, which connects sets of root morphemes with affixal morphemes. The morphonological type, therefore, determines which allomorphs can be appended to the root morpheme. Figure 13 shows the interface for "Root+Affix" morphotactic.

To fill in "Affix+Affix" morphotactics, a tabular representation of adjacency matrix for pairs of allomorphs is used, where rows contain allomorphs of one affixal morpheme, and columns contain allomorphs of another. The checkmark in this matrix means that the allomorph in column can follow the allomorph in row. Figure 14 shows the interface "Affix+Affix" morphotactics.

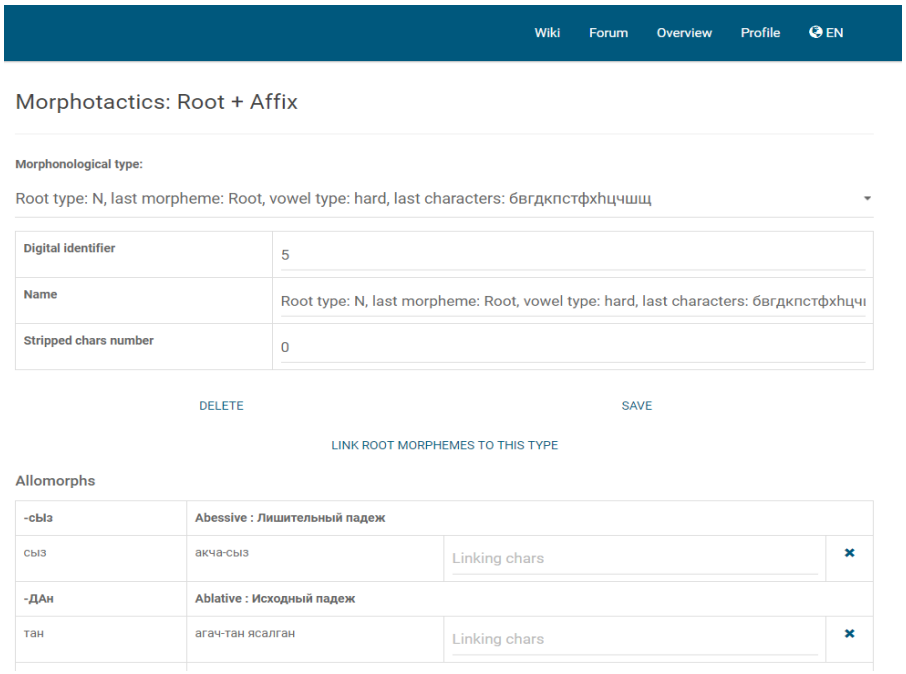


Fig. 13. "Root+Affix" view in expert mode

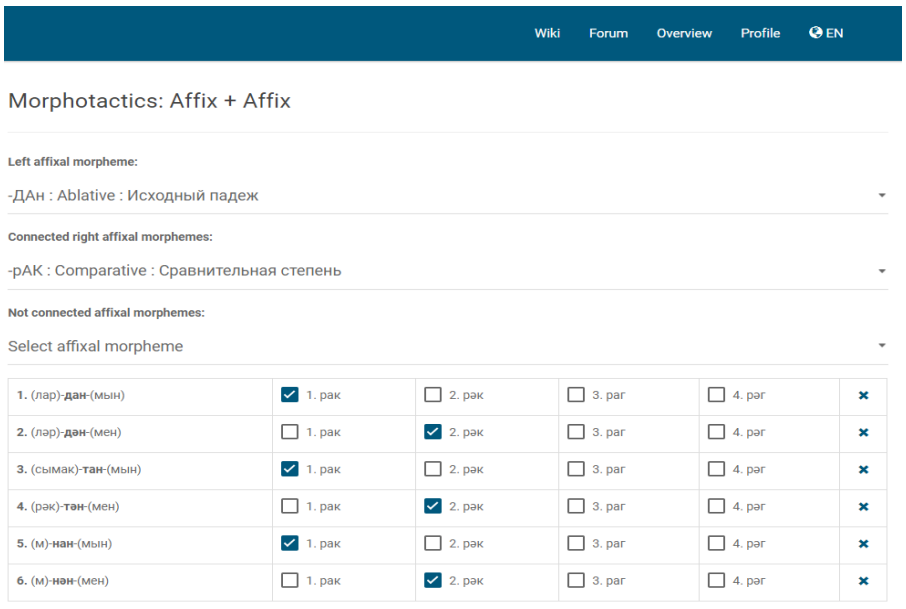


Fig. 14. "Affix+Affix" view in expert mode

4.4 Technical information

The server part of the portal is written in Python using the Django framework. The choice of language is explained by its simplicity, broad standard library and good support of packages for machine learning and natural language processing. This circumstance will make it possible in the future to facilitate the integration of other linguistic tools with the portal.

The Django framework, in turn, allows to effectively develop the standard web solutions, it automates the database interaction, GUI forms designing, and web-requests processing. The toolkit of this framework is made with taking into account the typical problems of web service development.

PostgreSQL was chosen as the database management system. The choice is justified on the one hand by the openness of this system, the presence of advanced functionality and high-grade optimization of queries. On the other hand, this DBMS has support of many external tools that make it possible to implement the requirements for extensibility and scalability of research software.

User interface is implemented separately in form of HTML pages with JavaScript code, generated on the server side using a template engine provided by the Django framework.

5 Results

The Turkic Morpheme Portal is at the stage of linguistic databases population with help of 30 experts in turkology and linguistics. The general statistics of the database are presented in Table 1.

Table 1. General statistics

Database entity	Count
Languages	37
Grammatical categories	19
Grammatical values	148
Concepts	15335
Affixal morphemes	780
Analytical morphemes	165
Root morphemes	35820
Multiword names	1618
Affixal morphotactics	21490
Morphonological types	47

At the time of paper preparation, the most complete are the databases for Tatar, Bashkir, Crimean Tatar, Kazakh, Uzbek, Kyrgyz languages. Table 2 contains detailed statistical information on databases for these languages.

Table 2. Statistics for the most complete languages

Database entity	Count
Tatar language	
Affixal morphemes	80
Analytical morphemes	18
Root morphemes	11896
Bashkir language	
Affixal morphemes	72
Analytical morphemes	0
Root morphemes	1256
Crimean tatar language	
Affixal morphemes	73
Analytical morphemes	0
Root morphemes	3672
Kazakh language	
Affixal morphemes	69
Analytical morphemes	1
Root morphemes	3938
Uzbek language	
Affixal morphemes	93
Analytical morphemes	78
Root morphemes	2116
Kyrgyz language	
Affixal morphemes	79
Analytical morphemes	53
Root morphemes	4175

6 Conclusion

The presented computer linguistic models and language processing technologies were considered in relation to Turkic languages, however, due to structural features, they are applicable to any agglutinative languages. Therefore, the formulated approaches to development of multipurpose multifunctional software based on the unified linguistic models are also applicable outside of context of turkology research.

Further development of the Turkic Morpheme Portal involves the development of new and integration of existing tools for Turkic languages processing on the basis of multilanguage model, which expands the possibilities for unification of linguistic software between supported languages.

The authors are confident that this portal will be actively used by the researchers of Turkic languages and developers of language processors, which will contribute to the creation and application of new common concepts in Turkic languages, especially in computer science and computer technology.

References

1. Guzev, V.G., Pyotrovski R.G., Sherbak A.M.: O sozdanii mashinnogo fonda tyurkskikh yazykov [About creation of machine fund for Turkic languages]. Sovetskaya tyurkologiya [Soviet turkology], 2, 92–101 (1988).
2. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: ICLR: Proceedings of the International Conference on Learning Representations Workshop Track, 1301–3781 (2013).
3. Fellbaum, C.: WordNet: an Electronic Lexical Database. MIT Press, Cambridge, MA (1998).
4. «Tugan Tel» Tatar National Corpus, http://tugantel.tatar, last accessed 2020/10/15.
5. Turkish National Corpus (TNC), https://www.tnc.org.tr, last accessed 2020/10/15.
6. Bashkir poetical corpus, http://web-corpora.net/bashcorpus/search/, last accessed 2020/10/15.
7. Turklang Electronic Corpora, http://www.turklang.net/en/resources-for-turkic-languages/, last accessed 2020/10/15.
8. Global Lexicostatistical Database, http://starling.rinet.ru/new100/mainr.htm, last accessed 2020/10/15.
9. Maps for Turkic Languages, http://turk.polycorpora.org, last accessed 2020/10/15.
10. Navigli, R., Ponzetto, S. P.: BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, 217–250. Elsevier (2012).
11. Palmer, M.: SemLink: Linking PropBank, VN and FrameNet. In: Proceedings of the Generative Lexicon Conference, GenLex-09, 13–17 (2009).
12. Plungyan, V.A.: Obshchaya morfologiya: Vvedenie v problematiku: Uchebnoe posobie. [General morphology: Problem introduction: Schoolbook]. Editorial URSS, Moscow (2000).
13. Gorodetsky, B.Y.: K probleme semanticheskoy tipologii [Onto semantic typology problem]. Moscow University Press, Moscow (1969).
14. Suleymanov, D.S.: Sistemy i informatsionnye tekhnologii obrabotki estestvenno-yazykovykh tekstov na osnove pragmaticheski-orientirovannykh lingvisticheskikh modeley [Systems and information tehcnologies of natural language processing on basis of pragmatically-oriented linguistic models]. Doctorate thesis on technical sciences, Kazan State University, Kazan (2000).
15. Suleymanov, D.S., Gatiatullin, A.R.: Strukturno-funktsionalnaya kompyuternaya model tatarskikh morfem [Structural and functional computer model of Tatar morphemes]. FEN Tatarstan Academy of Sciences, Kazan (2003).