Multilingual Turkic Thesaurus as Education Support Tool

Airat R. Gatiatullin[a], Nikolai A. Prokopyev[a], Madehur M. Ayupov[a], Djavdet S. Suleymanov[a] and Sofiya F. Mukhutdinova[b]

[a] *Institute of Applied Semiotics of Tatarstan Academy of Sciences, Kazan, 420111, Russia*
[b] *Kazan Federal University, Kazan, 420008, Russia*

Abstract

This paper presents integral linguistic model of the frame ontology and "Turkic Morpheme" linguistic portal, which is built on the basis of this model. The integrality of the model is based on fact that it combines two models of well-known linguistic resources namely WordNet and FrameNet. These models are expanded in the way of taking into account the structural and functional features of Turkic languages. The resources presented in the paper have two use cases for educational process. On the one hand, the portal database is a rich informational and reference material for the study of Turkic languages, and on the other hand, the ontology itself and the portal are used to teach semantic frames and ontologies in disciplines "Computational linguistics" and "Knowledge processing".

Keywords

Multilingual model, Educational linguistic resource, Turkology

1. Introduction

In the beginning of 1960s, there has been an increase in research in computer processing of Turkic languages, and in development of technologies for Turkic languages integration into digital domain. Despite this, all Turkic languages (except Turkish) are still considered low-resource languages.

One of the reasons for this, is that there is practically no real integration of Turkic languages processing research and linguistic resources creation for these languages. Developers duplicate linguistic models, resources, and software modules for NLP, which are basically 70-80 percent or more common to all Turkic languages. Now it is important to overcome such duplication, to join efforts on development of such resources and software modules.

This should increase the efficiency of multilingual text processing systems development and serve to solve other fundamental and applied tasks. To solve them, we need conceptual and formal linguistic models, databases that are common for Turkic languages, as well as software built on the basis of these models. Obviously, one of important requirements for such project is to present the results in form of public internet resources.

In 1988, the authors of article in the journal Soviet Turkology [1] presented a number of ideas for creating of unified electronic linguistic resources for Turkic languages, which were not implemented at that time. We will highlight two main ideas for our research:

1. To solve practical problems, it is necessary to create a large multi-level machine fund of Turkic languages (MMFTL), which should be built in such a way as to model both the common Turkic language structure and the structure of each specific language with all its inventory and structural units, rules of sign representation of language elements in speech, including the rules of linear deployment of *speech units.*

DTTL-2021: International Workshop on Digital Technologies for Teaching and Learning, March 22-28, 2021, Kazan, Russia

EMAIL: ayrat.gatiatullin@gmail.com (A.R. Gatiatullin); nikolai.prokopyev@gmail.com (N.A. Prokopyev); madehur@mail.ru (M.M. Ayupov); dvdt.slt@gmail.com (D.S. Suleymanov); sofiua-10@mail.ru (S.F. Mukhutdinova)
ORCID: 0000-0003-3063-8147 (A.R. Gatiatullin); 0000-0003-0066-7465 (N.A. Prokopyev); 0000-0001-9276-0098 (M.M. Ayupov); 0000-0003-1404-0372 (D.S. Suleymanov)

CEUR Workshop Proceedings (CEUR-WS.org)

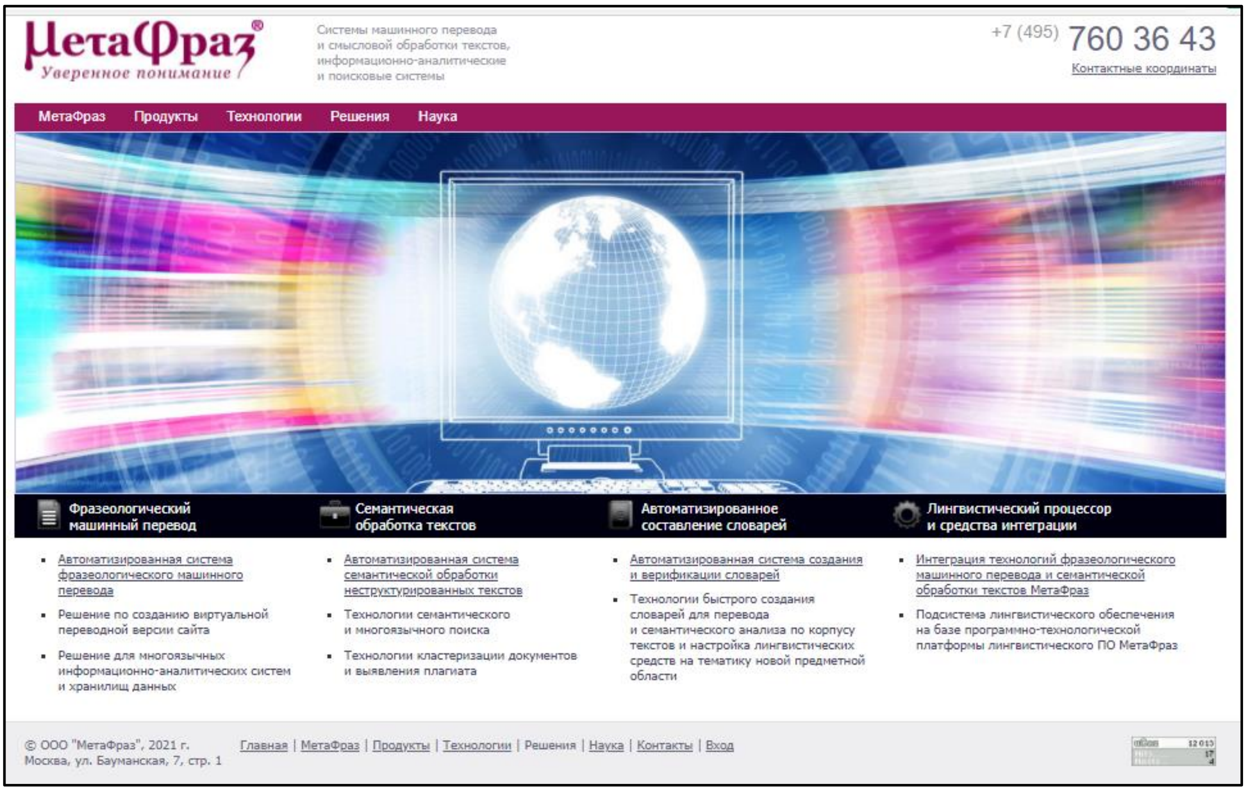
2. It is advisable to build MMFTL grammar block in form of a set of lexical and grammatical morphemes, as well as syntactic schema-models, subordinating the structure to the general architecture of the fund. These listed sets of morphemes and syntactic schemes for individual languages should be combined into a general Turkic grammatical thesaurus, which could be created using the MMFTL.

Since that time, the relevance of electronic resources for Turkic languages has not decreased. It is relevant despite the fact that there is currently a prevalence of technologies based on neural networks and machine learning in NLP. The main advantage of neural networks is that there is no need for detailed formalization of knowledge, since it is replaced by learning from samples. However, it is difficult to verbalize the results of neural network and explain why it made a particular decision. It is also impossible to guarantee repeatability and unambiguity of the results. This makes it difficult to use them in educational process.

Machine learning systems need large volume of data ready with morphological, semantic, and syntactic markup without ambiguity. And given the fact that almost all Turkic languages (except Turkish) belong to low-resource languages, there are no electronic corpora and software adequately suitable for such task. This is a sufficient justification for the claim that along with machine learning technologies development, work should continue on rule-based technologies, software, formal methods, and ontologies.

2. State of linguistic portals development

Currently, there are sufficient number of linguistic portals and platforms for widely spoken languages (for example, Russian and English). Most of these portals are intended for use in educational process, language learning in particular. Our analysis shows that the closest to the software we are developing in terms of provided resources and services is the Software-Linguistic Platform (SLP) “Metafraz” for Russian language (Figure 1). The platform description [2] indicates multilingual character of the platform, where this multilingualism is formed by Russian, English and German. “Metafraz” SLP is based on the theoretical concept of phraseological conceptual analysis of texts and provides the entire technological cycle of converting the document text into its formalized semantic representation.



Системы машинного перевода
и смысловой обработки текстов,
информационно-аналитические
и поисковые системы

+7 (495) 760 36 43

Фразеологический машинный перевод

- Автоматизированная система фразеологического машинного перевода
- Решение по созданию виртуальной переводной версии сайта
- Решение для многоязычных информационно-аналитических систем и хранилищ данных

Семантическая обработка текстов

- Автоматизированная система семантической обработки неструктурированных текстов
- Технологии семантического и многоязычного поиска
- Технологии кластеризации документов и выявления плагиата

Автоматизированное составление словарей

- Автоматизированная система создания и верификации словарей
- Технологии быстрого создания словарей для перевода и семантического анализа по корпусу текстов и настройка лингвистических средств на тематику новой предметной области

Лингвистический процессор и средства интеграции

- Интеграция технологий фразеологического машинного перевода и семантической обработки текстов МетаФраз
- Подсистема лингвистического обеспечения на базе программно-технологической платформы лингвистического ПО МетаФраз

© ООО "МетаФраз", 2021 г.
Москва, ул. Бауманская, 7, стр. 1

Figure 1: Software-Linguistic Platform “Metafraz”

This linguistic software is developed in form of a single integrated multifunctional software package consisting of several subsystems designed to solve separate processing tasks, formalization, translation and analysis of semantic content from multilingual documents. At the same time, the platform includes software modules that allows to create and adapt declarative tools for tuning to a given subject area while automating the creation of dictionaries from text corpus. “Metafraz” SLP includes the following subsystems:

- Text analysis subsystem.
- Subsystem for management and visualization of text processing.
- Text formalization subsystem
- Subsystem for declarative tools creation.
- Machine translation subsystem.
- Storage subsystem for declarative tools.

SLP provides the possibility of independent parallel data processing, distributed on different nodes of the Hadoop infrastructure, both at the level of message texts (i.e., each document is processed independently), and at the level of processing stages for a separate document (i.e., data extraction from one document is performed independently of data extraction from the same document).

3. Methodology

We propose an integral approach to the development of multilevel model for multilingual Turkic thesaurus. Among Russian research on the use of integral approach in computer linguistic models, we would highlight the “ETAP” project. During the project development, Y. D. Apresyan [3] described an integral approach, where the lexeme was proposed as the basic unit of model. In our model, taking into account structural and functional features of Turkic languages, we proposed to use the morpheme as the basic unit.

When creating the model in this research, we used the pragmatically-oriented approach to linguistic models, resources and software development, including a minimum set of tools sufficient to solve a certain range of problems. The pragmatic orientation in this project is to use a model as representation of agglutinative languages or, according to the classification of C. F. Hockett [4], set of languages with element-combinatorial morphology.

According to Hockett, the element-combinatorial model of morphology is a model focused on the “agglutinative standard” of word forms that allow for one-valued segmentation. The main approach of this model is linear segmentation. The model identifies allomorphs used in a particular context, as well as “non-separable morphemes”.

4. Integral linguistic model

We propose an integral pragmatically-oriented linguistic model for implementation of the multilingual Turkic thesaurus. Its integrality is achieved by combining several models of linguistic information representation, including frame ontology, proposed in [5] and linguistic ontology, proposed in [6].

The model of frame ontology proposed in [5] has the following notation:
$O_F = < C, R, S, G, T, D_S, D_G, E >$
$C = \{ c_i | 1, ..., n \}$ – a finite non-empty set of frame classes describing concepts of ontology domain
$R = \{ r_i | 1, ..., m \}$ – a finite set of binary relations on frame classes, $R \epsilon C \times C$, $R =\{ R_{ISA} \} \cup \{ R_{ASS} \}$
R_{ISA} – a set of set of antisymmetric, transitive, non-reflexive “class-subclass” hierarchical relations
R_{ASS} – a finite set of associative relations
$S = \{ s_i | 1, ..., k \}$ – a finite set of slots (class attributes)
$G = \{ g_i | 1, ..., l \}$ – a finite set of facets (slot attributes)
T – a finite non-empty set that defines a controlled dictionary of ontology domain terms
D_S – a finite set of slot types

D_G – a finite set of facet types
$E = \{ e_i | 1, ..., u \}$ – a finite set of class individuals.

The linguistic ontology proposed in [6] has the following notation:
$O = < L, C, F, G, H, R, A >$
$L = L_C \cup L_R$ – an ontology dictionary
L_C – a set of lexical units (signs) for concepts
L_R – a set of signs for relations
C – a set of ontology concepts
F – relations of language elements $\{ l_j \} \subset L_C$ with sets of concepts
G – relations of language elements $\{ l_j \} \subset L_R$ with sets of relations
$H \subset C \times C$ – taxonomic relations between concepts
R – non-taxonomic relations between concepts
A – a set of ontology axioms.

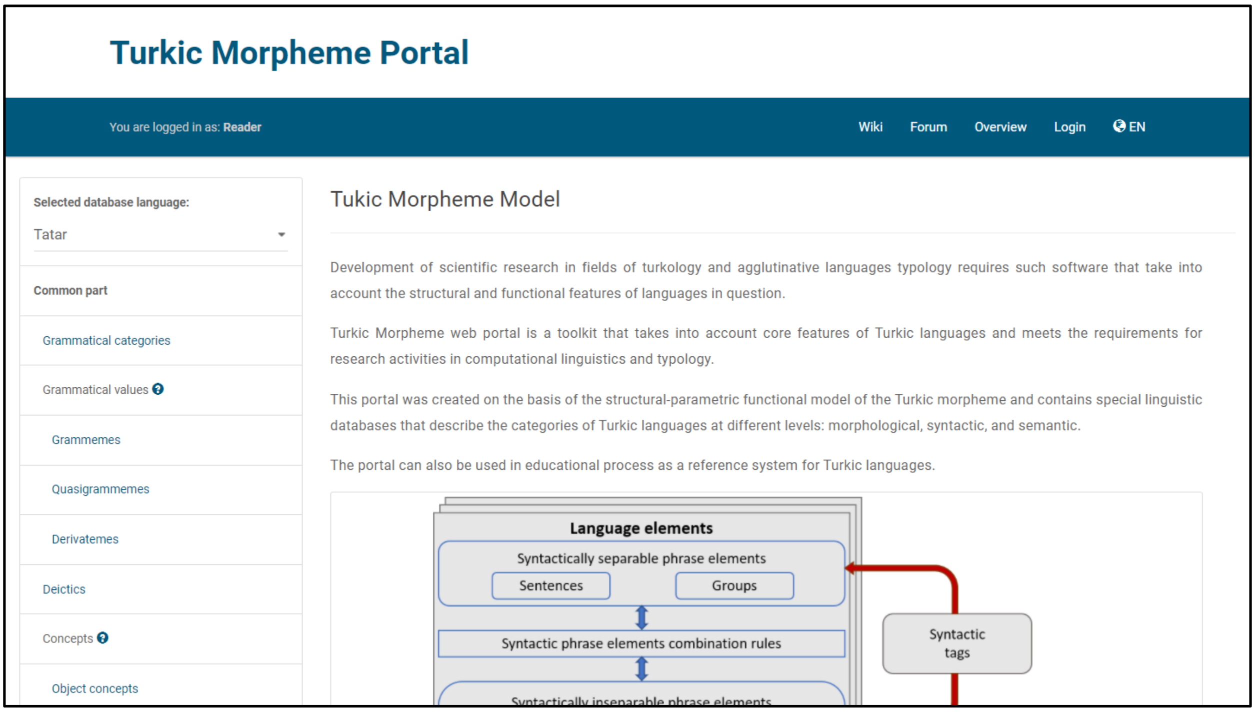
There are several projects which combine different types of ontological models. Among the most well-known resources created by combining electronic linguistic resources are BabelNet, Predict Matrix [7,8], SemLink [9]. They differ in the technologies used for combining: automatic or manual. What all these models have in common is that they combine existing resources. In contrast, our project assumes the creation of a new linguistic resource.

Based on the analysis of the previous ontological models, we propose the following ontological model:
$O_F = < C, R, S, L, H >$
C – a set of ontology concepts, in our model $C = C_O \cup C_{Act} \cup C_{Atr}$
C_O – object concepts
C_{Act} – action concepts
C_{Atr} – attribute concepts
$R = \{ r_i | 1, ..., m \}$ – a finite set of binary relations defined on classes, $R \epsilon C \times C, R =\{ R_{ISA} \} \cup \{ R_{ASS} \}$
R_{ISA} – a set of set of antisymmetric, transitive, non-reflexive “class-subclass” hierarchical relations
R_{ASS} – a finite set of associative relations, in our model these relationships are used to link roles
$S = \{ s_i | 1, ..., k \}$ – a finite set of slots for situation frames
$F = \{ f_i | 1, ..., l \}$ – a finite set of facets (slot attributes), in our model facets constitute role attributes which describe available concepts in role and signs for role notation
$L = L_C \cup L_R$ – ontology dictionary
L_C – a set of lexical units (signs) for concepts
L_R – a set of signs for relations
$H = H_C \cup H_R$
H_C – links between language elements and concepts
H_R – links between language elements and relations.

5. Software implementation

On the basis of the proposed ontology, a subsystem of the existing “Turkic Morpheme” portal [10] was developed, which implements the multilingual Turkic thesaurus. The “Turkic Morpheme” portal (Figure 2) is a set of services for computer processing of Turkic languages, which performs the following functions:

1. Information and reference educational system for Turkic languages;
2. Tools for database of the Turkic Morpheme Model;
3. Platform for communication of specialists in Turkology and Turkic languages processing;
4. Set of software modules that allow Turkic languages processing.



Tukic Morpheme Model

Development of scientific research in fields of turkology and agglutinative languages typology requires such software that take into account the structural and functional features of languages in question.

Turkic Morpheme web portal is a toolkit that takes into account core features of Turkic languages and meets the requirements for research activities in computational linguistics and typology.

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 portal can also be used in educational process as a reference system for Turkic languages.

Figure 2: Main page of “Turkic Morpheme” portal

One of the main portal functions is scientific and practical research in the field of Turkology. The multilingual Turkic thesaurus is naturally embedded in the integral model, key element of which is the Turkic morpheme. A morpheme is a minimal significant unit of a language, morphemes are divided into root and affixal types. The thesaurus complements functions of the portal itself with database entities and its own reference system, along with text processing services demanding semantic and syntactic data.

Server part of the portal is written in Python using the Django framework. The choice of language is determined by availability of many functions and software libraries for NLP. The Django framework, in turn, allows to effectively develop web applications, automates and simplifies the database interaction. The toolset of this framework already takes into account the most common tasks of web services development. PostgreSQL is used as a database management system (DBMS). This DBMS is open-source, supports advanced functionality and high level of query execution optimization.

User interface is implemented in form of HTML pages with JavaScript code generated on server using a special template engine provided by the Django framework. Accordingly, the developed portal subsystem for multilingual Turkic thesaurus uses all these technologies.

Within the framework of multilingual Turkic thesaurus subsystem, an information and reference educational service is implemented for the basic portal users (readers), a typologist's workspace for the common part of database interface, and an expert's workspace for the language-specific part of database interface. Therefore, there are three modes of subsystem usage: reader mode, typologist mode, and expert mode.

The common part of multilingual Turkic thesaurus contains information on semantic elements that are common to all languages included in the model. These elements are: concepts (objects, actions, attributes) and situation frames. Figure 3 shows the conceptual schema of database for the common part of thesaurus. Concepts express meanings with different relations between them. Situation frames consist of roles that are linked to a certain set of concepts in the context of particular situation.

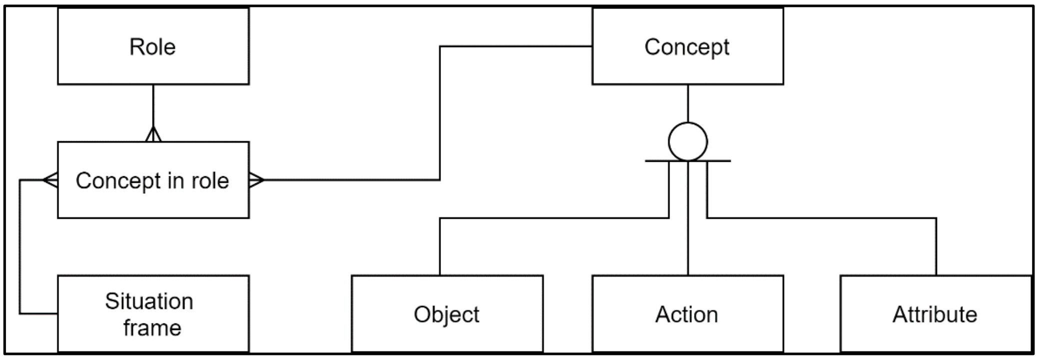
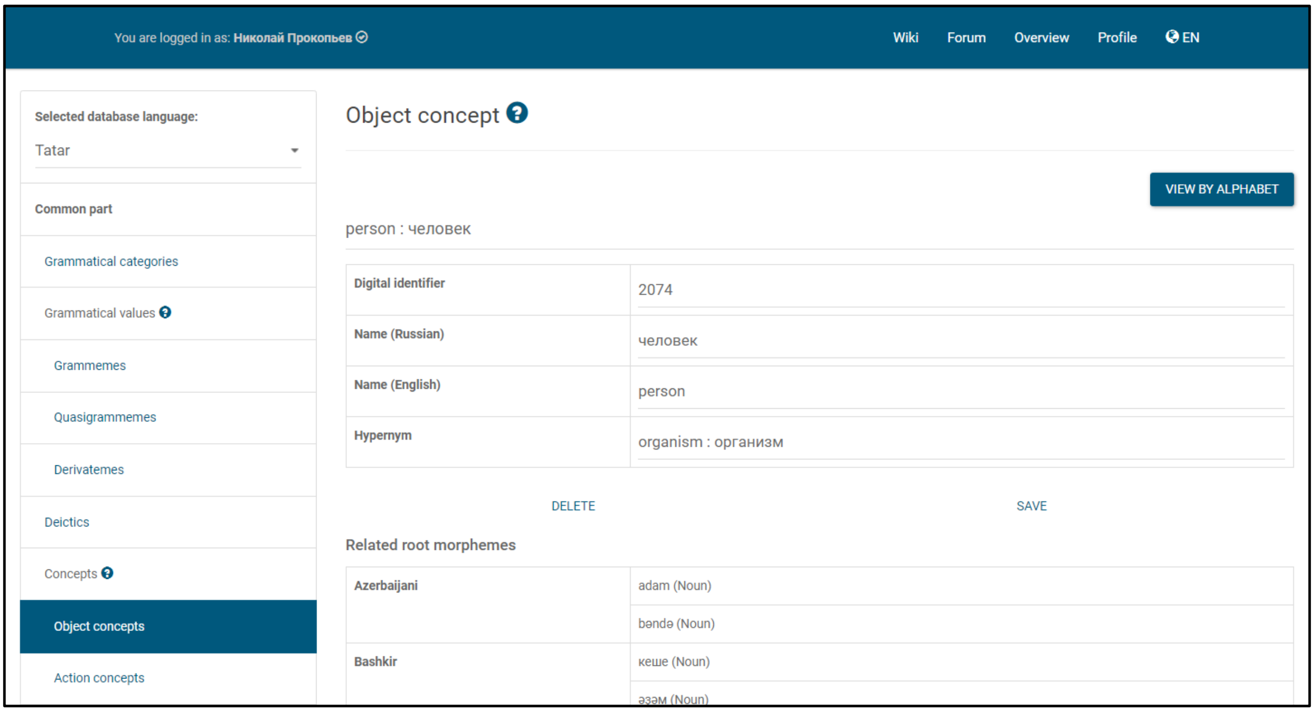


Figure 3: Conceptual ER-diagram of the common part of thesaurus

Database interface for the common part of thesaurus is available to both portal administrator and typologists through the functionality provided by the typologist's workspace (Figure 4). At the same time, the typologist has write-level access only to those concepts, frames and roles that he authored, while the administrator has access to every element of the common part.



Object concept

person : человек

Digital identifier	2074
Name (Russian)	человек
Name (English)	person
Hypernym	organism : организм

Related root morphemes

Azerbaijani	adam (Noun) bəndə (Noun)
Bashkir	кеше (Noun) әҙәм (Noun)

Figure 4: Typologist's workspace, concept editing form

The language-specific part of thesaurus includes those meaning representation elements that completely depend on the specifics of a particular language. They are root morphemes and language frames. Root morphemes are linked to concepts from the common part of thesaurus as some part of speech. Language frames implement situational frames for a specific language, using the grammatical part of Turkic Morpheme Model database, namely affixal and analytical morphemes and grammatical values. Figure 5 shows a conceptual schema of database for the language-specific part of the model.

The expert has full access to the language part for the corresponding language through the workspace functionality (Figure 6), can add, edit, delete root morphemes and language frames.

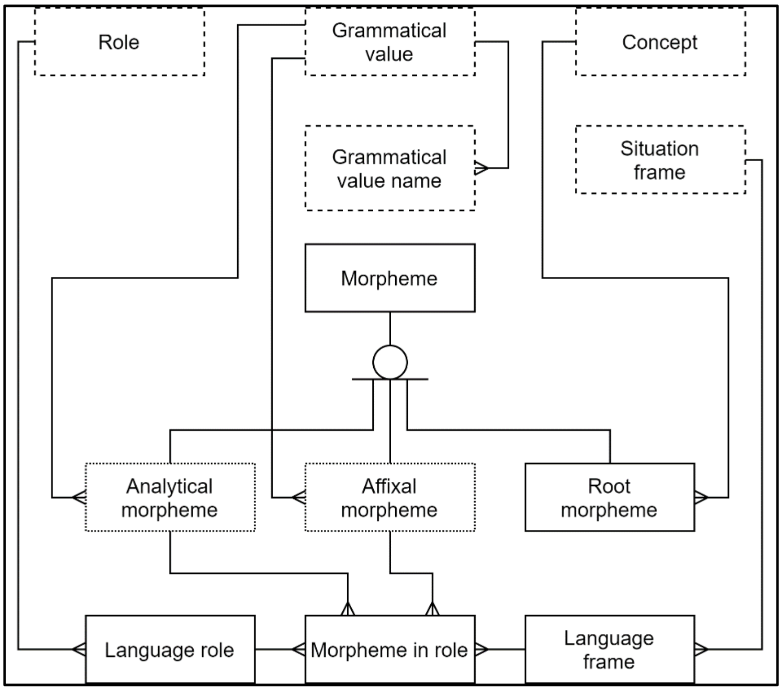
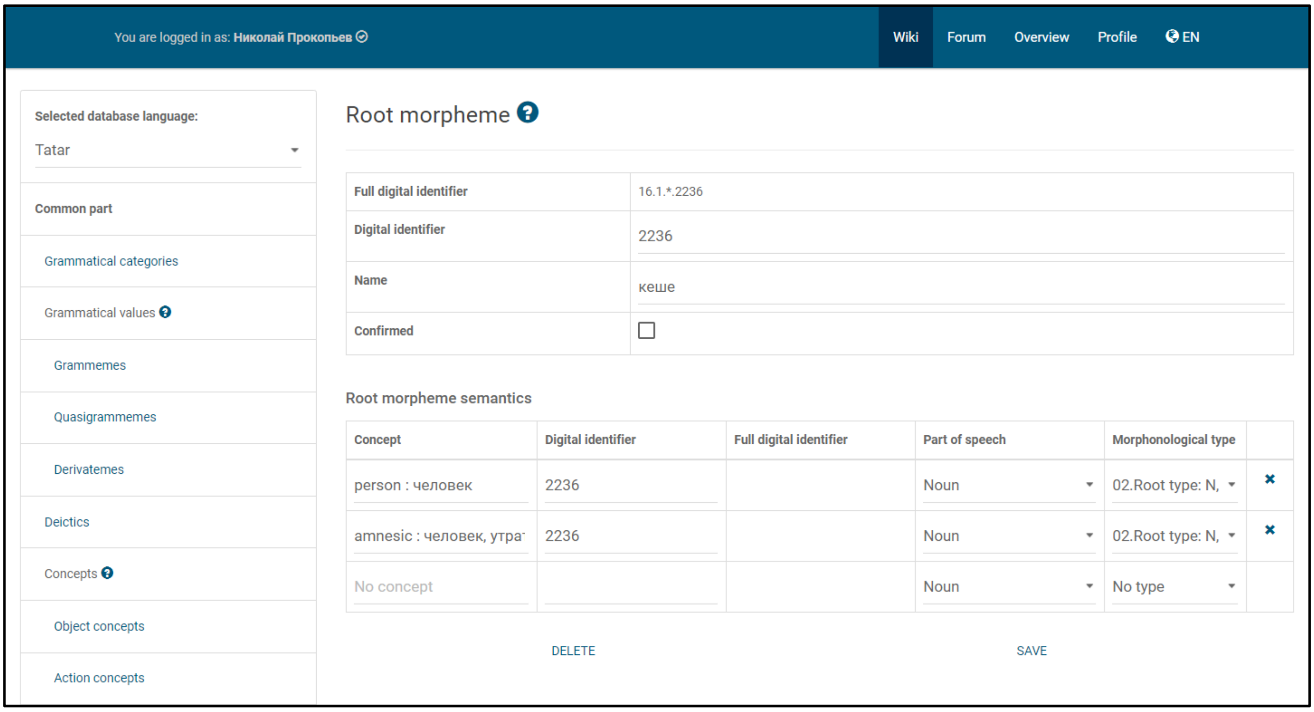


Figure 5: Conceptual ER-diagram of the language part of thesaurus



Selected database language: Tatar

Root morpheme

Full digital identifier	16.1.*.2236
Digital identifier	2236
Name	кеше
Confirmed	□

Root morpheme semantics

Concept	Digital identifier	Full digital identifier	Part of speech	Morphological type	
person : человек	2236		Noun	02.Root type: N,	×
amnesic : человек, утра	2236		Noun	02.Root type: N,	×
No concept			Noun	No type	

Figure 6: Expert's workspace, root morpheme editing form

6. Conclusion

The integral ontological model and the “Turkic morpheme” linguistic portal, presented in this paper, are already released and used by specialists both for accumulation of linguistic data on Turkic languages, and in educational process as a reference source for the study of Turkic languages or semantic frames and ontologies. Currently, the portal has more than 40 registered experts in various Turkic languages working on portal database. At the same time, the portal is constantly being developed, new modules and services are being implemented. In the future, it is planned to create an E-assessment system based on this portal, which will further expand its use in educational process.

7. References

[1] V. G. Guzev, R. G. Pyotrovski, A. M. Sherbak, O sozdanii mashinnogo fonda tyurkskikh yazykov [About creation of machine fund for Turkic languages], Sovetskaya tyurkologiya [Soviet turkology] No. 2 (1988) 92–101.

[2] Linguistic software “Metafraz R10”. URL: http://www.metafraz.ru/index/0-4.

[3] Y. D. Apresyan, Integral'noe opisanie yazyka i sistemnaya leksikografiya [Integral description of language and systemic lexicography]. Yazyki slavyanskoy kul'tury [Languages of Slavic culture], Moscow, Russia, 1995.

[4] C. F. Hockett, Two models of grammatical description, WORD Vol. 10 (1954) 210–234.

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

[6] A. Maedche, S. Staab, Learning Ontologies for the Semantic Web, in: Proceedings of Semantic Web Workshop, Hongkong, China, CEUR Workshop Proceedings 40 (2001) 1–10. URL: http://ceur-ws.org/Vol-40/maedche+staab.pdf.

[7] M. Lopez de Lacalle, E. Laparra, G. Rigau, Predicate Matrix: extending SemLink through WordNet mappings, in: Proceedings of LREC’14, Reykjavik, Iceland, 2014, pp. 903–909.

[8] M. Lopez de Lacalle, E. Laparra, I. Aldabe, G. Rigau, A Multilingual Predicate Matrix, in: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC’16), Portoroz, Slovenia, 2016, pp. 2662–2668.

[9] M. Palmer, (2009). Semlink: Linking PropBank, VerbNet and FrameNet, in: Proceedings of the Generative Lexicon Conference, Baltimore, MD, 2009, pp. 9–15.

[10] A. Gatiatullin, D. Suleymanov, N. Prokopyev, B. Khakimov, About Turkic Morpheme Portal, in: Proceedings of the Computational Models in Language and Speech Workshop, Kazan, Russia, CEUR Workshop Proceedings 2780 (2020), pp. 226–243. URL: http://ceur-ws.org/Vol-2780/paper19.pdf.