MULTIFUNCTIONAL MODEL OF MORPHEMES IN THE TURKIC GROUP LANGUAGES (ON THE EXAMPLE OF THE KAZAKH AND TATAR LANGUAGES)

D.Sh. Suleymanov
Scientific Research Institute of Applied Semiotics, Tatarstan Academy of Sciences Sciences
dvdt.slt@gmail.com,

A.R. Gatiatullin
Scientific Research Institute of Applied Semiotics, Tatarstan Academy of Sciences
ayratrg@antat.ru

A.B. Almenova
Scientific Research Institute of Applied Semiotics, Tatarstan Academy of
almen_akmaral-baijan@mail.ru

ABSTRACT

This article describes a multifunctional computer model of the Turkic affixal morphemes. This model is a hierarchical system of characteristics of morphemes belonging to different language levels: phonological, morphological, syntactic and semantic, and it requires a certain structure and unification in the description of characteristics of morphemes. It is a kind of "inventory" base of the language that can be used for different purposes; in particular, to perform automated comparative analysis of the properties of the Turkic languages, and to develop different linguoprocessors working with Turkic languages. Here, we describe the elements of the multifunctional computer model with examples on the Tatar and Kazakh languages.

Keywords: multifunctional model, affixal morpheme, Tatar language, Kazakh language.

1 Introduction

One of the problems in creating of linguoprocessors for Turkic languages is a deficit in structured data that would describe the properties of the Turkic language units. Obviously, the presence of such databases can accelerate the process of development of applied computer systems working with languages of the Turkic family, such as multilingual search Turkic corpora, machine translation systems for Turkic languages and others. Also, the presence of such models and database software, implemented as Web-interface, will be an effective assistant to turkologists in conducting different comparative studies.

A comparative analysis among Turkic languages with maximum automation of these processes requires conceptual models that would appropriately and adequately describe language units both structurally and functionally.

Another role of this model is to contribute to the process of unification of morphological categories, terms and tags, which are of particular importance for the representation of linguistic information. The analysis of the current situation shows that in the development of linguoprocessors working with Turkic languages and, in particular, in the Turkic corpus linguistics, despite the genetic, structural and typological commonness of Turkic languages, there is still lack of general principles and approaches to linguistic annotation of texts, to the system of tags for morphemes and morphological categories. In the future this may lead to difficulties in conducting comparative researches, as well as in the development of Turkic parallel corpora,

multilingual text processing systems and in resolving of other theoretical and applied problems. We suggest a multifunctional model that will help in the process of the comparative study of morphemes and categories, and also will serve as a unified information system on Turkic morphemes, available in the Internet space.

In the proposed multifunctional model, the morphemes of each of the Turkic languages are described both at the morphological level and at the level of other linguistic phenomena, such as phonology, syntax, semantics, and at the joint of language levels, such as the morpho-semantic and morpho-syntactic levels. The morpho-syntactic level studies the auxiliary particles and postpositions, which in some Turkic languages can be written as a single word, while in others should be written separately. The morpho-semantic level studies compound morphemes, such as mAgAe in the Tatar language, and their counterparts in other Turkic languages. To describe the semantic aspect of the multifunctional model, the authors develop a special unit in the form of situational frames [1].

Currently, the authors are developing a toolkit for filling of the proposed multifunctional computer model. At the same time, we are filling it on the example of the Tatar and Kazakh languages units. The basic elements of the model of morphemes for the Tatar language are represented in the work of the authors [2]. This article reveals the new results concerning the description of the semantic aspect of the model and mechanisms of expansion of the model as a unified framework for the description of language units of all the Turkic languages.

2 Model structure

The structure of this model is shown in table 1:

Table 1. Model structure

	Tatar	Kazakh	Turkish
Functio-nal aspect	Properties A[1].Tat	Properties A[1].Kaz	Properties A[1].Tur
Morpho-nological	Properties A[2].Tat	Properties A[2].Kaz	Properties A[2].Tur
Morpho-logical	Properties A[3].Tat	Properties A[3].Kaz	Properties A[3].Tur
Syntactic aspect	Properties A[4].Tat	Properties A[4].Kaz	Properties A[4].Tur
Semantic aspect	Properties A[5].Tat	Properties A[5].Kaz	Properties A[5].Tur

Every aspect is divided into sub-parameters, and those, in turn, can be further subdivided into sub-parameters, etc. [2].

In this article, we consider the fragments of the morphonological, functional and semantic aspects.

3 Morphonological aspect

The morphonological aspect is presented as a table of allomorphs. It describes all possible allomorphs of a morpheme, which constitute a superficial representation of the deep description of a corresponding morpheme in some context according to the phonological rules.

The rules are described through a set of production rules of the following aspect: A → B, where A is the context-condition of use of this allomorph, and B is the allomorph itself.
The context consists of the morphological and phonological components: A = A1 & A2

The morphological component determines what morpheme is on the left in the wordform, and the phonological one regulates what vowels and consonants are on the left in the wordform.

For example, in the Tatar language the morpheme -nIkI has 2 allomorphs: -nıkı, -neke. The selection of the necessary allomorph is determined by the vowel on the left (whether it is hard or soft), whereas in Kazakh there are 3 allomorphs: -niki, -diki, -tiki, and their choice is determined directly by the leftmost character.

4 Functional aspect

One of the parameters of the functional aspect is that of the coherence of the morpheme.

The coherence of the morpheme indicates whether the morpheme is free (analytical) or coherent (synthetic). This parameter is important for Turkic languages, because the morpheme can be written as a single word in one of the Turkic languages and separately in some other.

For example, the interrogative morpheme in the Tatar language is written as one word, whereas in the Kazakh language it is written separately. At the same time, in the Kazakh language it also has all the phonological variants depending on the context:
Tatar: -mI *urmanmı '(is it) a forest ?'*
Kazakh: MA *orman ba '(is it) a forest?'*

According to this criterion, a morpheme can be classified as an affix if it is coherent, or as an auxiliary part of speech if it is free.

Let us see an opposite example. The Tatar language has a postposition *belən* that is written separately from the main word, whereas in the Kazakh language the counterpart of this postposition is the morpheme *-Ben* with the allomorphs -ben, -men, -pen.

For example:
abıj belən 'with the brother'
aƶamen 'with the brother'

5 Semantic aspect

To describe the meaning of the linguistic units in the model, the authors propose to use relative-situational frames (RSF). RSF present the implementation of a typical situation, consisting of the name of the situation and a number of slots, which are the roles of the constructive elements of this situation. The slots of the frame are filled by language constructs called syntaxemes. The structure of syntaxemes is represented by morphemes.

RSF has the following representation:
SituationSi
Role1: SintaxemI1;
Role2: Sintaxem I2;
…
Role 3: Sintaxem IN;
End_Situation

The choice of a particular situation determines the choice of a particular type of RSF, which is called the base frame, with its corresponding slots-roles filled with some concrete values-syntaxemes.

For example, the basic RSF for the situation that expresses the action according to the change of state has the following representation:

Situation 7.2: action_statel
Object: Sintaxem 5;
Old_state: Sintaxem121;
New_state: Sintaxem119, 120;
Time: Sintaxem78;
Period: Sintaxem97;
Manner: Sintaxem99;
End_Situation

The authors conducted a classification of semantic contexts depending on types of relations that participate in the formation of the deep meaning of a given context. On the basis of this classification we obtained a system

consisting of 60 basic relational-situational frames.

Let us see an example of a syntaxeme.
Syntaxeme 118 as the value of a slot.

Direction:

1. Number of syntaxeme: 118
2. Main word: LSG (‘physical objects’)
    a. Syntax type: AF
    b. Morphological type: N
    c. Morphemic structure: -GA; -LAR-GA; -Im-GA; -Iŋ-GA; -sI-GA; -IbIz-GA; -IgIz-GA; -LAr-Im-GA; -LAr- Iŋ-GA; -LAr-IbIz-GA; -LAr-IgIz-GA
    d. The analytical form: taba
3. Main word: LSG (‘move’)
    a. Morphological type: V
    b. Morphological structure: *
4. Dependent word: -
    a. Morphological type: -
    b. Morphological structure: -
5. Meaning
    a. Type of situation: action_local
    b. The role of the syntaxeme: direction

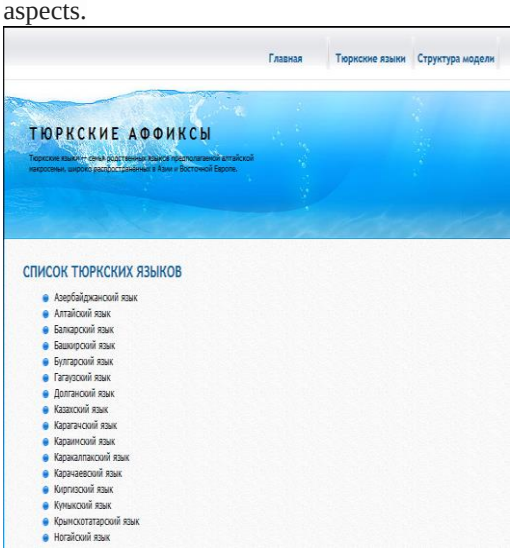
6 Programme description with pictures

The programme that is developed on the basis of the multifunctional model consists of the server and client parts. The server part is a relational database, and the client part is implemented as a Web-interface.

Let us consider the implementation of the interface elements in the form of a Web application. The programme is designed for different modes of use: the expert and the guest. It is expected that the experts will be granted access to the database for editing. The system administrator grants these rights after their registration. Other users can work with the programme as guests to make queries to the database for retrieval of the necessary information and to run the application programmes that work with the database.

The programme for working with the model is designed as multilingual, therefore on the first page of the website there is a list of languages of the Turkic group (Pic.1.). Thus, by selecting a language from this list, the user will be able to work with a single language or view the comparative information about all the Turkic languages.

The user can work with a specific aspect of the model. For this purpose, the main menu contains the option ‘Model Structure’. When chosen, it opens the list of aspects and sub-aspects.

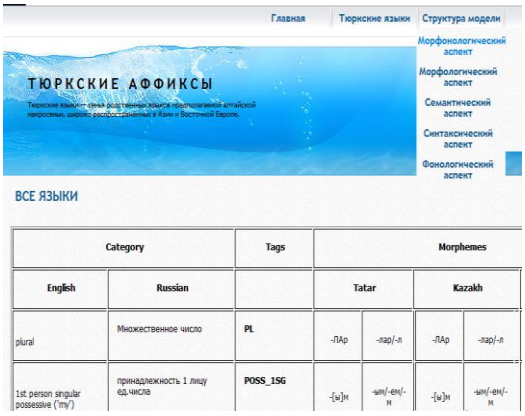


Pic.-1. “Home page” interface

As an example, let us see the interface of one the aspects – the morphonological one (Pic.2).

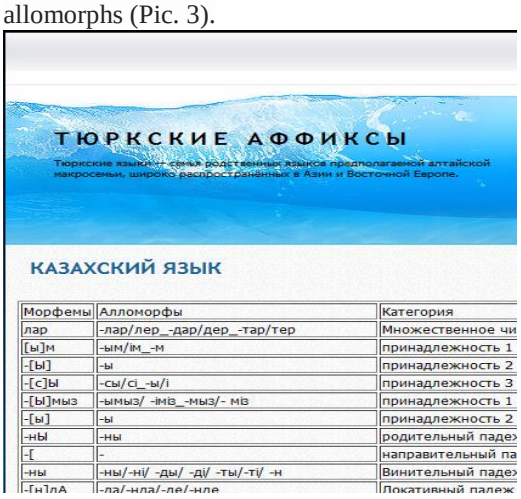
As can be seen in Pic.2, the table contains the information about the morphological categories, their designations in the form of grammatical tags, names of morphemes and lists of their allomorphs. In this picture a table

fragment with the information about the Tatar and Kazakh morphemes can be observed.



Pic.-2. Multilingual table of allomorphs

Morphonological information for each of the Turkic languages can be viewed separately. In this case, it has a detailed representation with an indication of the contexts of the use of allomorphs (Pic. 3).



Pic.-3. The morphonological aspect “The Kazakh language”.

It should be pointed out that the development of the programme is at an early stage and only a small number of functions are already implemented. The work on the programme is conducted simultaneously with the filling of the database model.

7 Conclusion

Due to the fact that the multifunctional computer model of morphemes described in this article is an open model, when being filled with concrete morphemes it can be supplemented by new aspects and new sub-parameters that characterize this morpheme. It is obvious that such characteristics as the adequacy of the model to the real linguistic phenomena or the completeness of the description of specific morphemes can be evaluated only on the basis of the stability of the model and its correct functioning in solving practical tasks on the basis of this model.

The model of a Turkic morpheme that is described in this article reflects the structure of morphemes, their purpose and their manifestation in the text. Apparently, it is a natural informative, educational and scientific base, as well as a database for building of different natural language processors.

8 References

[1] **SULEYMANOV, D.SH., GATIATULLIN, A.R.** *Strukturno-funktsionalnaya kompyuternaya model tatarskikh morfem* [Structural-functional computer model of the Tatar morphemes]. Kazan: FEN, 2003. 220 p. (rus)

[2] **SULEYMANOV, D.SH., GATIATULLIN, A.R.** *Napolneniye semanticheskikh slotov relyatsionno-situatsionnogo freyma na primere tatarskikh sintaksem* [Filling of semantic slots of the relational-situational frame on the example of the Tatar syntaxemes] // Collection of works of the conference “Open semantic technologies of designing intelligent systems” (OSTIS-2014). Minsk: BSUIR, 2014. P. 178. (rus)