EPiC Series in Language and Linguistics
Volume 1, 2016, Pages 321–330
CILC2016. 8th International
Conference on Corpus Linguistics

Building Corpus-Based Semantic Classifications of Tatar Ambiguous Affixes

Olga Nevzorova[1, 2], Alfiia Galieva[1], and Dzhavdet Suleymanov[1, 2]
[1] Tatarstan Academy of Sciences, Kazan, Russia
[2] Kazan Federal University, Kazan, Russia
onevzoro@gmail.com, amgalieva@gmail.com, dvdt.slt@gmail.com

Abstract

This study is aimed at exploring the semantic properties of Tatar affixes. Turkic languages have complicated morphology and syntax, which is a challenge for language processing.

The fundamental principle of inflection and derivation in Tatar, as well as in other Turkic languages, is agglutination, when the stem joins postpositive affixes in a strictly determined order.

The Tatar language has affixes of different types:

a) derivational affixes expressing only lexical meaning and forming new words;

b) inflectional affixes changing the word form (for example, case affixes);

c) affixes serving as means of derivation as well as inflection.

The current study is devoted to the ambiguous Tatar –lık polyfunctional affix which may be joined to nominal, adjectival and verbal stems and form derivatives of different types depending on contextual environment, the meaning of the stem and the composition of the affixal chain of a derivative. -Lık affix is a productive affix in modern Tatar which builds nominal, adjectival and verbal derivatives.

The answer to the question of the number of the types of derivatives and word forms produced with -lık affix is not trivial, and different researchers distinguish different types of derivatives.

Based on a thorough analysis of Tatar derivatives containing - lık affix we identified some empirical features of these constructs and then performed their manual and automatic classification. Four classes were distinguished. For our experiments we used data from the Tatar National Corpus “Tugan Tel” (http://corpus.antat.ru).

The results obtained may be used for disambiguation in Tatar National Corpus and for analyzing other Tatar ambiguous affixes.

A.Moreno Ortiz and C.Pérez-Hernández (eds.), CILC2016 (EPiC Series in Language and Linguistics, vol. 1), pp. 321–330

1 Introduction

The current project is implemented at the intersection of theoretical and computational linguistics in order to explore formal and semantic properties of Tatar linguistic units.

The rich agglutinative morphology and derivation of Turkic languages is a challenge for NLP. Available linguistic corpora of the Tatar language (see Section 2) give us reliable linguistic data containing linguistic units in their natural contexts, and this data require development of methodology of analysis. We consider the Tatar agglutinative morphology and the system of grammatical categories represented in grammatical annotation of the Tatar corpus as a key to semantic system of the language. Grammatical categories and meanings in languages are realized only in individual word forms, and morphological forms are special forms of interpretation of linguistic semantics that exist only within the lattice of grammatical phenomena. So selecting and combining grammatical, lexical and other parameters of corpus data, we may get certain sets of semantic samples.

The Tatar corpus has no system of semantic annotation yet, and corpus data is semantically unstructured, so extraction of semantic information from corpus data is not a trivial task. In our work we set the task of retrieving morphologically based semantic information, which requires detecting and describing the class of grammatically conditioned semantic phenomena. So a task of identifying these semantically class-specific cues and extracting semantic information with the help of morphological tags, syntactic structure and lexical co-occurrence of lexical items is a task of current importance.

Our study is aimed at exploring the semantic properties of Tatar ambiguous affixes, and this paper provides the first step into the automatic classification of Tatar word forms containing *-lık* affix within corpus data. *-Lık* affix is a productive affix in modern Tatar and it builds derivatives of different nature (nominal, adjectival and verbal ones).

In general, main stages of the work may be the following:

(a) extracting corpus contexts containing lexical items (word forms and derivatives) containing ambiguous affixes;

(b) study of contexts and manual semantic classification of these lexical items;

(c) development of rules for automatic classification and automatic classification of these items;

(d) automatic clusterization of contexts containing ambiguous affixes.

The paper is organised as follows: in Section 2 we outline some typological features of the Tatar language; in Section 3 we sketch the background of the research; Section 4 describes manually distinguished classes and criteria of classifying, Section 5 presents the results of automatic classification and evaluation, and Section 6 presents some conclusions and outlines directions for future research.

2 Tatar agglutinative morphology

Turkic languages have complicated morphology and syntax, which represents a challenge for language processing. The fundamental principle of inflection and derivation in Tatar, as well as in other Turkic languages, is agglutination, when the stem joins postpositive affixes in a strictly determined order. The stem of a word in Turkic languages has no indices of belonging to a certain class (for example, Turkic nouns are not characterised by gender, animacy or other suchlike features) and may be used without being overburdened with any morpheme.

Due to significant structural differences of Turkic languages from Indo-European languages, in the former there is no rigid distribution of words within parts of speech. The Turkic languages use a common morphological inventory for all parts of speech, and as a result there is no strict division

between nominal and verbal word forms. Nominal morphemes may be joined to verbal stems, forming various hybrid forms.

Example 1 represents joining the same Directive case affix *-ga* to the nominal and verbal stem:

(1) *Min ul bazar-ga bargan-ga şatlanam.*
I he market-DIR go-PAST_IND, DIR be glad
I am pleased that you are going to the market.

The Tatar language has affixes of different types:
a) derivational affixes, expressing only lexical meaning and forming new words;
b) inflectional affixes changing the word form (for example, case affixes);
c) affixes serving as means of derivation as well as inflection.
The last type is called the *polyfunctional* affix in Tatar linguistics [Tatar grammar, 1993].

Depending on the meaning of the stem and the character of the affix chain, these affixes may form new words (2) or express a grammatical meaning (3) without changing the word class of its stem.

(2) *başsız malay*
head- ABESS boy
'a stupid boy'

(3) *başsız jaydak*
head- ABESS rider
'a headless rider'

Having multiple different functions in the language system, such affixes are a challenge both to theoretical and computational linguistics, being highly ambiguous. Nevertheless, word forms and derivatives formed by means of polyfunctional affixes have differing structure, and they occur in particular collocations and have restrictions in use that enables us to describe them in terms of such groupings of properties.

In derivatives and word forms, affixal chains have strict rules of successive addition of elements, depending on the type of the word form.

Another significant feature of the Tatar language is that there are no inflectional paradigms in traditional sense, as closed sets of word forms (for example the set of all inflected forms based on a single stem). A paradigm in Turkic languages may have fuzzy borders, because the number of potential elements that may be added (inflectional affixes) is indefinite.

The current study is devoted to the ambiguous Tatar *–lık* morpheme which may be joined to nominal, adjectival and verbal stems and form derivatives of different types depending on contextual environment, the meaning of the stem and the composition of the affixal chain of a derivative. The empirical data for classification experiments were derived from the Tatar National Corpus (http:// corpus.antat.ru).

3 Related work

Nominal semantic classes and derivatives in languages of different types share properties that make them significant for disclosure of a number of linguistic phenomena. Grammatical categories and meanings in languages are realized only in individual word forms, and grammatical categories are special forms of interpretation of linguistic semantics. So by selecting and combining grammatical, lexical and other parameters of a query, we may get certain sets of semantic samples.

In particular, (Bell et al., 2012) present cue-based noun classification for English and Spanish. The main objective of the work is to automatically acquire lexical semantic information by classifying nouns into previously known noun lexical classes. This is achieved by using particular aspects of linguistic contexts as cues that identify a specific lexical class (Bell et al., 2012).

There has been very little research in the automatic semantic classification of Turkic affixes. The rich agglutinative morphology and derivation of Turkic languages is a challenge for NLP. Most of the challenges stem from the complex morphology and how morphology interacts with syntax. (Oflazer, 2014).

There are some special works devoted to ambiguous affixes in Turkic languages, and almost all of them are devoted to theoretical issues. In special literature different types of Turkic ambiguous affixes are described. E. Sevortyan described the history of formation and development of Turkic ambiguous affixes (Sevortyan, 1952). Azerbaijani researcher A. Gasanov identified types of affixes depending on their functioning and distinguished lexical, grammatical, lexical-grammatical and gramatical-lexical ambiguous affixes (Gasanov, 1980). Homonymous affixes in Turkic languages are also studied by M. Dzhurabayeva (Dzhurabayeva, 1975), M. Akhtyamov (Akhtyamov, 1966) and other researchers.

F.A. Ganiev analysed some aspects of correlation of lexical and grammatical meanings in suffixes of modern Tatar (Ganiev, 2005) and gave a brief description of polyfunctional affixes.

Ambiguity of affixes is a matter of great difficulty and a challenge for Turkic language processing (Oflazer, & Kuruöz, 1994; Hakkani-Tür, Oflazer, &Tür, 2002; Daybelge &. Çiçekli, 2007; Oflazer, 2014).

Tatar grammars contain some information on polyfunctional affixs (Tatar Grammar, 1993), and ambiguous Tatar affixes are considered by F. Ganiev (Ganiev, 2005), and R. Salakhova (Salakhova, 2007).

There have been no special studies of derivatives and word forms, containing Tatar ambiguous affixes, implemented on corpus data, representing real distribution of words and word forms.

By now two linguistic corpora for the Tatar language have been developed:

the Tatar National Corpus (http://corpus.antat.ru);

the Corpus of Written Tatar (http://corpus.tatar/en).

These corpora provide for a plenitude of representation of a wide range of linguistic phenomena and ensure typicality of data. Our research is based on linguistic data from the Tatar National Corpus.

«Tugan Tel» Tatar National Corpus is a linguistic resource of the modern literary Tatar language. The project is carried out within the framework of the "Preservation, study and development of the official languages of the Republic of Tatarstan and other languages in the Republic of Tatarstan for 2014-2020" State Program.

The volume of the Corpus is 82,000,000 word forms (by the end of 2015). The Corpus contains texts of different styles and genres (fiction, media texts, official documents, educational and scientific literature, etc.).

All the texts included in the Tatar Corpus go through special procedures of meta-annotation (attributing metadata to the text). The Corpus has a system of grammatical annotation that is oriented at presenting all the existing grammatical word-forms. Grammatical annotation of a Tatar word includes the information about its part of speech and a set of morphological features (parameters). Morphological annotating of Corpus texts is carried out using the module of two-level morphological analysis of the Tatar language implemented with the help of the program tool PC-KIMMO (Suleymanov et al., 2013).

The search system of the Corpus enables us a search for lexemes, word forms and individual grammatical parameters. The system of grammatical annotation is based on Leipzig Glossing Rules (a number of tags, specific for the Tatar language, has been added).

The Tatar language has a set of ambiguous affixes. Depending on the meaning of the stem and the character and structure of the affixal chain, these affixes may form new words or express a grammatical meaning without changing the word class of its stem. Having multiple different functions

in the language system, ambiguous affixes are a challenge both to theoretical and computational linguistics.

As an object of research, we chose - *lık* affix that is a productive affix in modern Tatar which builds mainly nominal, adjectival and verbal derivatives. It has four allomorphs *-lık /-lek, -lıg, /-leg,* the choice of which rigidly depends on the phonetic features of the stem and the structure of the affixal chain of a linguistic item.

The answer to the question of the number of types of derivatives and word forms produced with *–lık* morpheme is not trivial, and different researchers distinguish different types of derivatives.

For example, F. Ganiev considers 7 basic derivational and 2 grammatical meanings of the *-lık* morpheme (Ganiev, 2005).

R. Salakhova (Salakhova, 2007) considers 6 types of nominal derivatives, one type of adjectival and one type of verbal derivatives.

In (Ganiev 2005) and (Salakhova, 2007) the semantic classes of derivatives are very fractional and the criteria for referring derivatives to a certain class are rather vague.

All the researchers emphasize that the *–lık* morpheme is very productive and is in active use.

The objective of the automatic semantic classification of Tatar derivatives containing a certain morpheme has not been posed yet, although the problem of the automatic semantic classification of words of different types in different languages is being tackled by many researchers.

4 Semantic classification of -lık derivatives

Based on a thorough analysis of Tatar derivatives containing *-lık* affix we identified some empirical features of these constructs and determined their main grammatical and semantic features and then performed manually a semantic classification of *-lık* derivatives.

The object of classification are word forms and derivatives containing *-lık* affix; the linguistic data is taken from the Tatar National Corpus. 28,000 contexts (full sentences) containing -lık items were extracted from Corpus data in course of our work, and four classes were distinguished.

Table 1 below provides a description of these classes.

The work completed shows that semantic classes of Tatar linguistic items containing *-lık* affix are characterized by different structure and different contextual environment, so these characteristics may be generalized and rules for each class may be defined.

5 Automatic classification of -lık forms and evaluation of results

Basing on the analysis of Tatar derivatives containing *-lık* morpheme from a theoretical point of view, we identified some empirical features of these constructs and performed their automatic classification.

The object of classification embrace word forms and derivatives containing *-lık* affix; the linguistic data is taken from the Tatar National Corpus. As far as the Corpus is not disambiguated, in its contexts *-lık* morpheme may be marked by means of two different tags: NMLZ (nominalizer) and PSBL (possibilitive form). Altogether the Corpus contains 207,000 word forms and derivatives with *-lık* affix, 28,000 contexts were extracted.

Attribute	Class 1	Class 2	Class 4	Class 4
IsAttached	noun, adjective, numeral other stems	noun stems	verbal stems	verbal stems
HasFunction	derivational affix	links words within a sentence	derivational affix	serves as a means for attaching a subordinate clause to the main clause
HasMeaning	nouns of broad meaning (concrete nouns, abstract nouns and collective nouns)	forms attributive words denoting a measure of something (including time measure). A typical meaning of derivatives of class 2: enough for sth.'	forms attributive words A typical meaning is potentiality, possibility, ability (inability in the negative form) to accomplish an action	makes sentential arguments
Examples	(1) *Ak* 'white' *+lık; aklık* 'whiteness' (2) *narat* 'pine-tree' *+lık* ; narat*lık* 'pinery' (3) t*aş* 'stone, rock' *+lık ; taş-lık* 'stony ground' (4) *un 'ten' +lık ; unlık* 'ten-rouble note'	(1) *Ber atna-lık azık* One week-NMLZ food Food for a week (2) *50 bit-lek kitap* 50 page-NMLZ book A book of 50 pages (for example, for publishing) (3) *ber külmäklek tukıma* one dress-NMLZ fabric a piece of fabric suffice for a dress	(1) *tab-ar-lık azık* find-FUT_IND, PSBL food 'food that one may find' (2) *bir-er-lek äyber* give-FUT_IND, PSBL thing 'a thing that one may give (to somebody)'	(1) *Kar **yaw-gan-lık-tan cir** ak.* Snow fall- PCP_PS, NMLZ, ABL ground white. The ground is white because snow has fallen. (2) *Keşelär minem kemgä **kayt-kan-lıg-ım** turında bähäsläşä.* Man-PL my who-DIR return-PCP_PS, NMLZ, POSS_1 about dispute_PRES People dispute to whom I have come.

Table 1: Semantic classes of word forms and derivatives containing *-lık* affix

We identified some empirical features of *-lık* constructs and built relevant rules for semantic classification and performed an automatic classification of them. Four classes of derivatives containing *-lık* affix were distinguished.

Rules for automatic semantic classification are based on:

- type of stem (noun, adjective, verb, etc.);
- affixal structure of the word form containing *-lık*;

- left context of the word form containing -lık (part of speech tags of the two elements to the left)
- right context of the word form containing -lık (part of speech tags of the two elements to the right);
- exception to the rules for each class.

Table 2 represents general scheme of rules (without unnecessary details) for automatic semantic classification.

Class	**Type of stem**	**Essential elements of the structure of a word form**	**Left context**	**Right context**	**Exception to the rules (examples of lemmas)**
Class 1	noun adjective numeral	absence of participle affixes	irrelevant feature	irrelevant feature	Yawlık 'shawl'
Class 2	noun	attributive affixes ATTR_MUN or ATTR_ABES	numeral or number	noun	Hujalık 'farm, household'
Class 3	verb	affix of future participle (PCP_PS) before *-lık*	irrelevant feature	noun	Batırlık 'courage', sabırırlık 'patience'
Class 4	verb	affix of past participle (PCP_FUT or PCP_FUT_NEG) before *-lık,* and case or/and possessive affix after *-lık*	irrelevant feature	irrelevant feature	Tuganlık 'kinship'

Table 2: General scheme of rules for automatic classification.

Stem Type	Number of *-lık* derivatives
N	121,000
ADJ	112,000
V	34,000
NUM	4,000

Table 3: Distribution of *-lık* derivatives depending on types of stem.

In this section, we will also present some results of statistical analysis of contexts containing *-lık* derivatives concerning distribution of these contexts. As we mentioned above, the Tatar language tends to have universal means for expressing meanings from the same conceptual grammatical domain and tends to have the same affixes for different parts of speech. Table 3 represents distribution of *-lık* derivatives depending on types of stem.

Table 4 demonstrates distribution of semantic classes of *-lık* derivatives (results of the automatic classification).

Class 1	**Class 2**	**Class 3**	**Class 4**	**Other**	**Total**
18,881	3,550	4,655	1,290	276	28,000

Table 4: Distribution of semantic classes of *-lık* derivatives

The same context may contain derivatives of different types - linguistic items belonging to the same semantic class or belonging to different classes. In particular, analysis of contexts containing 5 or more *-lık* constructs shows that in 82% of contexts they are derivatives related to the same class, and an overwhelming majority of them functions as homogeneous parts of the sentence.

Table 5 demonstrates the distribution of contexts containing *-lık* affix.

Number of *-lık* affixes in the context	**Number of *-lık* derivatives in the same context**
1	162,000
2 or more	29,000
3 or more	5,000
4 or more	998
5 or more	347

Table 5: Distribution of contexts with *-lık* affix

We evaluated results of automatic classification (Table 6).

Class	**Precision**
Class 1	92%
Class 2	95%
Class 3	99%
Class 4	99%

Table 6: Evaluation of results of automatic classification

The worst results were obtained for Class 1. This is easily explained by the fact that according to the developed classification rules, if a word is not related to any class it is to be related to class 1.

6 Conclusions and Future Work

An important task is to develop a methodology for semantic classification of linguistic items relying upon corpus data. Taking into account that Tatar corpus has not a system of semantic annotation, a task to extract semantic information relying upon grammatical structure and contextual environment of linguistic items seems promising.

This paper provides the first step into the automatic classification of Tatar word forms containing ambiguous *-lık* affix within corpus data. First we examined the corpus contexts containing units with *--lık* affix and developed rules for classifying these units, with four classes distinguished. Then we built rules for automatic semantic classification of derivatives and word forms containing *-lık* affix, accomplished automatic classification of Tatar corpus contexts, and evaluated the results.

The results of our study improve the results of preceding theoretical studies of Tatar polyfunctional affixes accomplished on limited linguistic data that was arbitrary, limited and contingent in many respects. The thesis about the multifunctionality of the *-lık* affix was confirmed on corpus data, relevant semantic classes are distinguished.

The results of our study also may be used for corpus disambiguation. As we mentioned above, currently *-lık* affix is marked by means of two different tags (NMLZ and PSBL).

The work carried out clearly indicates that

- tag MNLZ is to be attributed to classes 1, 2, 4;
- tag PSBL is to be attributed to class 3.

A task of identifying class-specific cues based on morphological, syntactic and lexical co-occurrence features provide indicative hints for retrieval particular classes of semantic phenomena from unstructured corpus data. Gained experience may be used for analysis of other Tatar and Turkic ambiguous affixes. The developed methodology also may be used for automatic classification of particular meanings of Tatar indirect cases (especially Directive, Ablative and Locative) on corpus data.

Another intended direction of the future work is automatic clusterisation of corpus contexts containing *-lık* affix in order to verify our results or to have other classes demonstrating distribution of formal and semantic properties of *-lık* contexts.

Acknowledgements

The reported study was funded by RFBR according to the research project № 15-07-09214a.

References

Akhtyamov, M. (1966). Formation of Lexical-Grammatical Homonyms in Bashkir (the Bashkir vocabulary), Ufa.

Baldwin, T. (2005). General-purpose lexical acquisition: Procedures, questions and results. In *Proc. of the Pacific Association for Computational Linguistics 2005*, Tokyo, Japan, pp. 23-32.

Baldwin, T. & Bond, F. (2003). Learning the countability of English nouns from corpus data. In *Proc. of the 41st Annual Meeting of the ACL*, Sapporo, Japan, pp. 463-470.

Bel N, Romeo L, Padró M. Automatic Lexical Semantic Classification of Nouns. In *Calzolari N, Choukri K, Declerck T (et al.), editors. Proc. of the Eight International Conference on Language Resources and Evaluation (LREC'12);* 2012 May 23-25; Istanbul, Turkey. Paris: European Language Resources Association; 2012. p. 1448-1455.

Corpus of Written Tatar http://corpus.tatar/en (accessed April 4, 2016)

Ganiev, F. A. (1974). *Suffixal derivation in modern Tatar language*, Kazan. (in Russian).

Ganiev F.A. (2005). Types of Suffixes in Turkic Languages. *Scientific Tatarstan*, 2005, 1-2:107-111. (In Russian).

Gasanov, A.A. (1980) *Homonymy in the Azerbaijan language*. Baku.

Daybelge, T., .Çiçekli I. (2007). A Rule-Based Morphological Disambiguator for Turkish. In *Proc. of Recent Advances in Natural Language Processing* , pp. 145-149.

Dzhurabayeva, M. (1975). *Affixal Homonymy in Uzbek*, Tashkent.

Hakkani-Tür, D. Z., Oflazer K. & Tür G. (2002). Statistical morphological disambiguation for agglutinative languages. *Computers and the Humanities*, 36:4. Pp. 381-410.

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proc. of the 28th Annual Meeting on the ACL.* Association for Computational Linguistics. Pp. 268-275.

Kenesei, I. (2014). On a multifunctional derivational morpheme: Its use in relational adjectives or nominal modification, and phrasal affixation in Hungari. *Word Structure;* Oct 2014, Vol. 7 Issue 2. http://www.nytud.hu/kenesei/publ/MultifunctAfxHung2013.pdf

Oflazer, K. (2014). Turkish and its Challenges for Language Processing. *Lang Resources & Evaluation* (2014) 48:639-653. DOI 10.1007/s0:79-014-9267-2

Oflazer, K., &Kuruöz, Ì. (1994).Tagging and morphological disambiguation of Turkish text. In *Proc. of the fourth conference on Applied natural language processing*. Association for Computational Linguistics. Pp.144-149.

Salakhova, R.R. (2007). *Homonymous Suffixes of the Tatar Language*. Kazan. (in Russian)

Sevortyan, E. (1952). On the relation of grammar and vocabulary in the Turkic languages. *Theory and history of the language in the light of the works of V. Stalin on linguistics.* Moscow. Pp. 306-367.

Suleymanov, D., Nevzorova, O., Gatiatullin, A., Gilmullin, R., &,Khakimov, B. (2013). National corpus of the Tatar language "Tugan Tel": Grammatical Annotation and Implementation. *Procedia-Social and Behavioral Sciences*, *95*, pp. 68-74.

Tatar Grammar: in 3 volumes. (1993) Kazan: Tatar Publishing Company, V. 1.

"Tugan Tel" Tatar National Corpus //http://corpus.antat.ru (accessed April 4, 2016)

Yüret, D., Türe, F. (2006). Learning Morphological Disambiguation Rules for Turkish. In *Proc. of HLT-NAACL'06*, 2006, pp. 328-334.