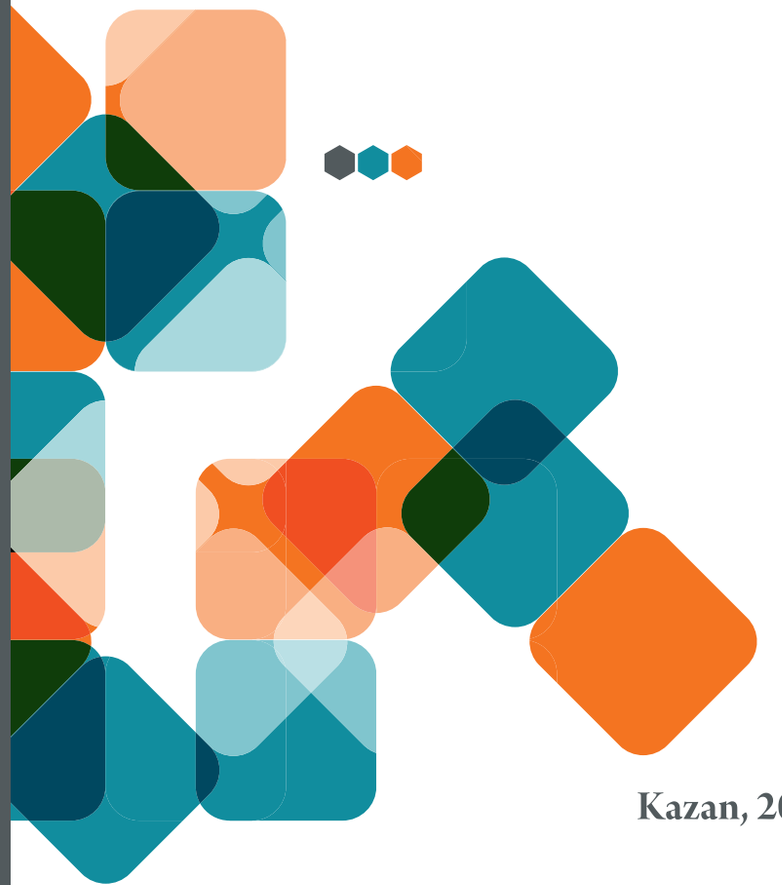
PROCEEDINGS
OF THE INTERNATIONAL CONFERENCE
"TURKIC LANGUAGES PROCESSING"

TurkLang 2015

September 17-19, 2015, Kazan, Tatarstan, Russia



Kazan, 2015

TurkLang • 2015

Tatarstan Academy of Sciences

PROCEEDINGS
OF THE INTERNATIONAL CONFERENCE
"TURKIC LANGUAGES PROCESSING"

TurkLang-2015

September 17–19, 2015, Kazan, Tatarstan, Russia

Kazan
2015

UDC 002.001.4
BBC 32.973.8122
P 93

Tatarstan Academy of Sciences
Research Institute of Applied Semiotics
L.N. Gumilyov Eurasian National University
Ministry of Education and Science of the Republic of Kazakhstan
Scientific Research Institute of Artificial Intelligence
Russian Foundation of Basic Research
Russian Association of Artificial Intelligence
Kazan Federal University
Institute of Philology and Intercultural Communication
Institute of Computational Mathematics and InformationTechnologies
The Higher Institute for Information Technology and Information Systems
"Selet" Tatarstan Republic Youth Social Fund

The publication was supported by RFBR, research project No. 15-46-07007.

Printed by decision of the Editorial Board of the Tatarstan Academy of Sciences

P 93 **Proceedings of the International Conference "Turkic Languages Processing: TurkLang-2015".** – Kazan: Academy of Sciences of the Republic of Tatarstan Press, 2015. – 488 c.
ISBN 978-5-9690-0262-3

These proceedings include papers presented at the International Conference on Turkic languages processing "Turklang-2015" (Kazan, Tatarstan, Russia, 17–19 September 2015). The Conference is focused on the relevant problems of computational linguistics in Turkic languages. The participants discussed issues related to the development of formal linguistic models, corpora projects, machine translation tasks, applied systems and technologies of computer and cognitive linguistics. These proceedings were designed for researchers, teachers and students specializing in the field of computer and cognitive linguistics and its applications.

UDC 002.001.4
BBC 32.973.8122

ISBN 978-5-9690-0262-3

FOREWORD

These Proceedings include papers presented at the International Conference on Turkic languages Processing "Turklang-2015" (Kazan, Tatarstan, Russia, 17–19 September 2015).

These Proceedings were published with financial support of the Russian Foundation for Basic Research, project №15-46-07007.

The participants of the Conference were scientists and specialists from Russia (Kazan, Moscow, Bashkortostan, Yakutia, Chuvashia, Tuva, the Crimea, and others), Azerbaijan, Kazakhstan, China, Kyrgyzstan, Turkey, Uzbekistan, the United States and the Czech Republic. The Conference is focused on the relevant problems of computational linguistics in Turkic languages. The participants discussed issues related to the development of formal linguistic models, corpora projects, machine translation tasks, applied systems and technologies of computer and cognitive linguistics. Common features in the lexis, morphology, syntax and semantics of Turkic languages allow researchers to use similar approaches, methods and technologies in their projects.

The subject of the Conference is in constant development. Today, it includes a new area focused on unification of grammatical annotation systems in the corpora of Turkic languages that was thoroughly discussed within the Uniturk seminar ("Unification of Grammatical Annotation Systems in the Electronic Corpora of Turkic Languages"). Currently, there is a lack of a single unified annotation system for Turkic languages, including standard tags for morphemes and morphological categories. Unification of corpora annotation systems is not a trivial practical task and it requires theoretical reconsideration of many traditional grammatical descriptions.

The creation of new terminology in Turkic languages is an important issue. The appendix to these Proceedings contains a new terminological dictionary on computer science for four languages (English-Russian-Tatar-Chuvash Dictionary of Computer Terms).

The organizers of the Conference would like to thank the Director of the Institute of Computational Mathematics and Information Technologies of Kazan Federal University (KFU) R. H. Latypov, the Director of the Institute of Philology and Intercultural Communication of KFU R. R. Zamaletdinov, the Director of the Higher Institute for Information Technology and Information Systems of KFU A. F. Khasianov, as well as members of the Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences for their contribution to the organization and success of the "Turklang-2015" Conference.

D. Sh. Suleymanov
Chairman of the "Turklang 2015" Program Committee

PROGRAM COMMITTEE:

Dzhavdet Suleymanov (Kazan, Tatarstan, Russia) – Chairman
Altynbek Sharipbayev (Astana, Kazakhstan) – Co-chairman
Aelita Salchak (Kyzyl, Tuva, Russia)
Anna Dybo (Moscow, Russia)
Ayrat Hasyanov (Kazan, Tatarstan, Russia)
Eshref Adaly (Istanbul, Turkey)
Gavril Torotoev (Yakutsk, Saha, Russia)
Gulila Altenbek (Urumqi, China)
Kemal Oflazer (Doha, Qatar)
Lenara Kubedinova (Simferopol, Crim, Russia)
Masuma Mamedova (Baku, Azerbaijan)
Radif Zamaletdinov (Kazan, Tatarstan, Russia)
Rustam Latypov (Kazan, Tatarstan, Russia)
Sergei Tatevosov (Moscow, Russia)
Tashpolot Sadykov (Bishkek, Kyrgyzstan)
Ualisher Tukeev (Almaty, Kazakhstan)
Valerian Zheltov (Cheboksary, Chuvashiya, Russia)
Zinnur Sirazitdinov (Ufa, Bashkortostan, Russia)

ORGANIZING COMMITTEE:

Olga Nevzorova (Chair), (Kazan, Tatarstan, Russia)
Ayrat Gatiatullin (Scientific Secretary), (Kazan, Tatarstan, Russia)
Madekhur Ayupov (Kazan, Tatarstan, Russia)
Alfiya Galieva (Kazan, Tatarstan, Russia)
Ramil Gataullin (Kazan, Tatarstan, Russia)
Rinat Gilmullin (Kazan, Tatarstan, Russia)
Aidar Khusainov (Kazan, Tatarstan, Russia)
Bulat Khakimov (Kazan, Tatarstan, Russia)
Marat Kurmanbakiev (Kazan, Tatarstan, Russia)

SECTION 1 | **MACHINE TRANSLATION TECHNOLOGIES**

STUDY OF THE PROBLEM OF CREATING STRUCTURAL TRANSFER RULES AND LEXICAL SELECTION FOR THE KAZAKH-RUSSIAN MACHINE TRANSLATION SYSTEM ON APERTIUM PLATFORM

Abduali Balzhan[1], Akhmadieva Zhadyra[2], Zholdybekova Saule[3], Tukeyev Ualsher[4], Rakhimova Diana[5]

[1] KazNU named after Al-Farabi, Almaty, Kazakhstan
balzhan_5696@mail.ru
[2] KazNU named after Al-Farabi, Almaty, Kazakhstan
[3] KazNU named after Al-Farabi, Almaty, Kazakhstan
[4] KazNU named after Al-Farabi, Almaty, Kazakhstan
[5] KazNU named after Al-Farabi, Almaty, Kazakhstan

Active integration of Kazakhstan into the world community and the increasing volume of information flow between our country and its foreign partners, and a real need of different segments of population for operational machine translation while using the Internet, determine the relevance of machine translation between the Kazakh language and various major world languages, like English, Russian, French, German, and recently, Chinese languages, as well as in the vice versa machine translation. The priorities of information interaction for the population of Kazakhstan with foreign partners and internally are mainly defined by interaction in three languages: Kazakh, English and Russian. In this regard, it is highly relevant to have highly efficient instrumental support machine translation for the trilingual language interaction. So are actual research and development industrial quality machine translation systems from Russian language to Kazakh language, and vice versa. Analysis of the state of research in the field of machine translation from Russian into

Kazakh shows that research in this area is practically nonexistent, despite the presence of two or three commercial machine translation software products, the quality of the translation which is not high enough. We create Kazakh-Russian translation system with using Kazakh lexical rules from English-Kazakh and we based on the Russian-Tatar Apertium platform. And we create Kazakh-Russian dictionary on the Apertium platform. We search and make some rules for this language pairs.

1. Introduction

Automation and improvement of translation quality is very actual problem in the sphere of artificial intelligence. As we know, the organization of machine translation – a set of interrelated stages performing algorithms. In the field of machine translation by the main topical issue there is a problem of quality of machine translation. So far various methods of machine translation are developed from one natural language to another.

In this article describes a problem of creating structural transfer rules for sentences and lexical selection for the Kazakh-Russian and Russian-Kazakh language pairs on a platform Apertium.

2. Structural transfer rules

Three type of dictionaries are used in Apertium platform for lexical processing: monolingual dictionaries, for morphological analysis and generation of Russian, Kazakh and bilingual dictionaries for Kazakh-Russian, Russian-Kazakh, lexical transfer.

In the Kazakh-Russian dictionary, apertium-kaz-rus.kaz-rus.dix is filled with words and their translations. For example:

```
“<dictionary>
 <alphabet></alphabet>
 <sdefs>
  <sdef n=”num” c=”Имя числительное”/>
      …
 </sdefs>
<pardefs>
<pardef n=”__num_gender”>
```

<e> <l></l><r><s n=”m”/><s n=”an”/><s n=”sg”/><s n=”nom”/></r> </e>
<e> <l></l><r><s n=”m”/><s n=”an”/><s n=”sg”/><s n=”det”/></r> </e>
<e> <l></l><r><s n=”m”/><s n=”an”/><s n=”sg”/><s n=”ord”/></r> </e>
…
</pardef>
<e> <l>бір<s n=”num”/></l><r>один<s n=”num”/></r> <par n=”__num_gender”/></e>”.

We create for words new paradigm for numerals. It is for do not write one analyses for all words. In this paradigm we write gender, case, number.

And for Adjectives we create same paradigm with numerals. Adjectives has three degrees of comparison.

<pardef n=”__adj_sint”>
<e> <l></l><r><s n=”m”/><s n=”an”/><s n=”sg”/><s n=”nom”/></r> </e>
<e> <l></l><r><s n=”m”/><s n=”an”/><s n=”sg”/><s n=”det”/></r> </e>
<e> <l></l><r><s n=”m”/><s n=”an”/><s n=”sg”/><s n=”ord”/></r> </e>
<e> <l><s n=”subst”/></l><r><s n=”m”/><s n=”an”/></r> </e>
<e> <l><s n=”comp”/></l><r><s n=”comp”/></r> </e>
</pardef>

Then in the period of translating some words, which have two meaning, it can be seen that sometimes words in not applied part-of-speech tag right. For example “сорока человека”. There is word “сорока” has two meaning: 1. Number – “forty” and 2. View of bird – “magpie”. To solve this problem we must write rules for this situation. And in the apertium-rus.rus.rlx we write rule:

Number: for “Сорок человек” – genitive
SELECT Gen IF (0 Num) (1 N + Gen) ;

To improve quality of translation it is very important to fill dictionary with words with correct part of speech tags.

3. Lexical selection

All words in a sentence related in meaning. The machine translators in translating an ambiguous word in many cases, do not translate correctly. To solve this problem you need to use the rules of lexical selection. Lexical selection – selection of the respective translation of the original proposal. [3]

Template used in the lexical selection.

<Rule> – the beginning of the rules;

<Match lemma = "specified keywords"> – defines the word;

tags = "parts of speech" – speech tag of the defining words, for example, a noun – "n", name prilogatelnoe – "adj", t.s.s.;

<Select lemma = "Choose Your Word" – the choice of the respective transfer "defines the word";

tags = "parts of speech" – a tag that indicates the part of speech the word translated treated;

</ Match>, </ rule> – closing to appropriate tags.

These lexical rules are in the open / free code platform Apertium, the module apertium-kaz-rus.kaz-rus.lrx.

<rule>
<match lemma="көру" ><select lemma="смотреть"/></match>
</rule>
<rule>
<match lemma="түс" tags="n"/>
<match lemma="көру" tags="v.*">
<select lemma="видить" tags="*.perf.*"/></match>
</rule>

As an example, the Kazakh word "көру" translated into Russian as " видеть, смотреть." If the sentence " Мен түс көрдім " word " көрдім " combined with " түс" is on the lexical rule translated as " видеть ", and in other cases, translated as " смотреть ".

Currently we considered methods and sampling and the lexical grammar of the variable is in their machine translators.

4. Results

Running Kazakh-Russian (and vice versa) systems translate simple phrases and sentences. In Kazakh-Russian bilingual dictionary contains 9043 word.

5. Conclusion

As a result of the solution of these tasks were developed bilingual and monolingual dictionaries on a platform Apertium, also were investigated structural transfer rules for sentences and the rules of a lexical selection, was executed experimental check and assessment of machine translation.

REFERENCES

Печерских, Т. Ф., Амангельдина, Г. А. (2012) “Особенности перевода разносистемных языков (на примере английского и казахского языков)”, Молодой ученый. №3, 259–261 [http://www.moluch.ru/archive/38/4406/];

Documentation on a wide variety of development and usage scenarios can be found on the Apertium Wiki (http://wiki.apertium.org/);

http://beta.visl.sdu.dk/constraint_grammar.html.

CHOOSING THE MODEL FOR SOLVING THE PROBLEM OF LEXICAL SELECTION FOR ENGLISH-KAZAKH LANGUAGE PAIR IN THE FREE/OPEN-SOURCE PLATFORM APERTIUM

Dina Amirova

Al-Farabi Kazakh National University, Information Systems Department,
Al-Farabi av., 71, 050040, Almaty, Kazakhstan
amirovatdina@gmail.com

This paper describes rule-based lexical selection for English-Kazakh pair in the free/open-source platform Apertium and describes the model for lexical selection that can be applied to Kazakh language as a target language. The problem of lexical selection is one of the main tasks of word processing, which is associated with the task of word-sense disambiguation. It will be not difficult to choose the correct meaning of words to people, but for machines it isn't simple. Despite the long history of its existence, a word sense disambiguation still is the developing branch of the knowledge. The machine translation system Apertium consists of several modules. One of them is the lexical module which is considered there. This lexical module in the translation ambiguous word of input language to the target language selected one lexical form of all possible with the help of rules depending on the context. Rules are hand-written. There given examples of rules that are used to selection of right sense of ambiguous words. Also to solve this problem can be used statistical models. In the paper would be considered a statistical model, maximum entropy model, which is used for solving a problem of lexical selection. Maximum entropy model shows high accuracy in different systems. The use of two systems, rule-based selection and statistical-based selection, for solving the problem of lexical selection can give a more accurate translation of texts. In the paper will be considered the works which are done to this time for solving the problem of lexical selection.

1. Introduction

Recently, the role of the Republic of Kazakhstan in the international arena increases, which leads to a significant increase of interest in our country from the world community. Today English language is recognized as international language. The official language of the Republic of Kazakhstan is Kazakh. The scope of work of translators is increasing every year. Accordingly, the creation and developing of

automated translation from English into Kazakh is very important and useful for people who want to translate to Kazakh.

One of the main tasks of processing of texts is the task of a lexical selection which is connected to the task of a word sense disambiguation. It is the correct choice of the word or term in accordance with the context in which they are used. Solving the task of ambiguity is one of the central word processing task. Word-sense disambiguation is used in different areas: to improve the quality of machine translation, improve the accuracy of methods of classification and clustering texts, information retrieval and other applications.

Today, there are many algorithms and models of resolving it. Linguists distinguish next kind of ambiguity: lexical, morphological, syntactic, Let's consider lexical ambiguity. Lexical selection is a choosing one translation of the word in target language by context of source language. Lexical selection is a main task of processing language.

2. Rule-based lexical selection in Apertium platform

Apertium is a platform of machine translation which development started with financing from the governments of Spain and Catalonia at Alicante University (Universitat d'Alacant). It is a free software which is published by developers according to GNU GPL conditions. To create the new system of machine translation one needs develop linguistic data (dictionaries, rules) in accurately specified XML formats [1]. The Apertium machine translation system consists following modules: deformatter, morphological analyser, part-of-speech (POS) tagger, lexical transfer, lexical selection, structural transfer, morphological generator, post-generator [2].

In rule-based free/open source platform Apertium [3] the problem of lexical selection is solved by module of lexical selection (F.M. Tyers, M.L. Forcada 2013). The rules are written by hand. The rules of lexical selection are written a way in which translation is taken by depending located near words. Hand-written rules do not always cover the entire context. So to solve this problem we use statistics methods and models, which connected with training corpora to generate rules automatically.

Rule-based lexical selection is written in file apertium-eng-kaz.eng-kaz.lrx for language pairs from English into Kazakh. This lexical module in the translation ambiguous word input language to the target language

selects one lexical form of all possible using rules which depend on the context. All the rules are written in the XML-format.

The content of the lexical rules:
<rule> – start of rule;
<match lemma=”the word in english/kazakh” defining word;
tags=”part of speech” tag of the words part of speech,
for example, noun – “n”, adjective – “adj”, and etc.;
<select lemma=”selected word” selection of a particular ambiguous word translation;
tags=”part of speech” tag of the words part of speech;
</match>,
</rule> – closing of the relevant tags.

Example of lexical selection rule for ’zhas’:

```
<rule>
<match lemma=”year” tags=”n.pl”>
<select lemma=”” tags=”n.*”/>
</match>
</rule>
<rule>
<match lemma=”year” tags=”n.pl”>
<select lemma=”” tags=”n.*”/>
</match>
<match lemma=”old” tags=”adj.*”/>
</rule>
```

(Example from apertium-eng-kaz.eng-kaz.lrx)

3. Statistical-based lexical selection

Statistical-based lexical selection chooses the most likely translation with their probability. Statistical-based lexical selection based on counting frequency of collocation or words in corpora. One of the main part of statistical machine translation system is to make corpora especially corpora of large volume. One of the difficult task is a collection of parallel corpora, in our case, to gather the corpus of Kazakh and the corpus of the English. Now we have been developing a bilingual corpus, which contains 4255 sentences. Corpora are collected from fairytales, books. Today we are training these corpora. To receive

corpora-based lexical selection we need aligned corpora, which is not easy to do. Kazakh language has a complex morphology. So, some words can be aligned to several words.

Because of rule-based lexical selection do not cover all cases of ambiguity, we want to use both of type of lexical selection, which were meant above. For creating statistical-based lexical selection we collect and develop bilingual corpus. Then we are training system by adding words to monolingual dictionary of Kazakh (apertium- kaz.kaz. lexc) and English (apertium-eng-kaz.eng.dix) language and adding to bilingual dictionary (apertium-eng-kaz.kaz-eng.dix).

Maximum Entropy Model

Today there are different types of models and methods of solving the problem of lexical ambiguity, which are used to solve it. One of this model's Maximum Entropy model [4].

Model maximum entropy lexical selection includes a set of binary functions and appropriate weights for each function. The feature is defined as $h^s(t,c)^2$, where t is a translation, and c – is a source language context.

$$h^s(t,c)^2 = \begin{cases} 1, & if\ value\ t\ under\ condition\ c \\ 0, & in\ other\ cases \end{cases}$$

During the learning process each function is assigned a weight λ^s, and combining the weights as in Equation (2) gives the probability of a translation t for word s in context c.

$$p_s(t|c) = \frac{1}{z} exp \sum_{k=1}^{n_F} \lambda_k^s h_k^s(t,c), \quad (2)$$

where Z – is a normalizing constant. Thus, the most probable translation can be found using equation (3)

$$\hat{t} = \underset{t \in T_s s}{argmax}\ p_s(t|c) = \underset{t \in T_s s}{argmax} \sum_{k=1}^{n_F} \lambda_k^s h_k^s(t,c) \quad (3)$$

It is important to note that the rules for the feature $h^s(t,c)^2$ will be different depending on the language pair.

4. Results

Today there are 85 rules in English-Kazakh lexical selection. The system can translate simple phrases and sentences with ambiguity [5].

5. Conclusion

In this paper was described a lexical module for English-Kazakh language pair in the free/open-source platform Apertium, where lexical selection problem is solved by writing rules for words. In the future we would like to use maximum entropy model for more effective solving the problem of lexical selection. Because this model shows the high accuracy. We are preparing parallel corpora for English and Kazakh languages now, which are collected from fairy-tales and books. Then we would train it as statistical machine translation system has a property of «self-learning». As a method would be used supervised learning. This method based on word-alignment from corpora.

REFERENCES

1. Apertium: http://en.wikipedia.org/wiki/Apertium

2. Sundetova, A., Karibayeva A., Tukeyev Ua.: STRUCTURAL TRANSFER RULES FOR KAZAKH-TO-ENGLISH MACHINE TRANSLATION IN THE FREE/OPEN-SOURCE PLATFORM APERTIUM. Proceedings of the International Conference on Computer processing of Turkic Languages ”TURKLANG’14”, Istanbul(2014)

3. The Apertium machine translation platform: http://apertium.org/

4. Francis Morton Tyers. Feasible lexical selection for rule-based machine translation. Ph.D. thesis. – Universitat d’Alacant. – May, 2013.

5. Сундетова А. М., АПЕРТИУМ ПЛАТФОРМАСЫНДАҒЫ АҒЫЛШЫН-ҚАЗАҚ МАШИНАЛЫҚ АУДАРМА ЛЕКСИКАЛЫҚ МОДУЛІ. Международная научная конференция студентов и молодых ученых «Фараби әлемі». – Алматы: «Қазақ университеті», 2014. – С. 145.

THE ONTOLOGICAL MODEL OF NOUN FOR KAZAKH-TURKISH MACHINE TRANSLATION SYSTEM

Lena Zhetkenbay[1], Altynbek Sharipbay,
Gulmira Bekmanova, Unzila Kamanur

L.N. Gumilyov Eurasian National University, Astana, 010000, Kazakhstan
[1]jetlen_7@mail.ru

In this work, we discussed the structure of ontological model according to noun in Kazakh and Turkish languages for machine translation system. This model gives us an opportunity to compare generally the similarities and differences of the languages. There can be used informational searching, making machine translation and autoreport, dialogical and other systems.

1. Introduction

The ontological models of nouns for Kazakh and Turkish machine translation system are built in this research work.

Kazakh language belongs to Kipchak group [1] whereas Turkish language belongs to Oghuz group [2] of Turkic languages. The agglutination is one of the peculiarities of Turkic languages characterized by a large number of word types for each word formed by adding affixes to its end (suffixes and endings). There exists a strict order for adding of affixes: suffixes are the first to be attached to a stem or root of the word, then the endings of plural number, possessive endings, case endings and conjunction endings. All these features of Turkic languages assume a slight formalization for morphological and syntax rules as far as they have a strict order of adding affixes from a morphological point of view and a strict word order in sentence from the point of view of syntax. Therefore the issue of machine translation can give decentish results on the basis of grammar rules.

At present the types of machine translation systems are varied. The choice in use of the machine translation system depends on the complexity of the formalization of natural language or national linguistic corpus of natural language.

The works [3–9] are devoted to the issues of machine translation from Turkish language into other Turkic languages. The issues of formalization of grammar rules in Kazakh language are solved in the

works [10–21]. All these results will substantively facilitate the creation of Kazakh-Turkish machine translation system.

2. The ontological models of nouns for Kazakh and Turkish machine translation system

Ontology is a powerful and widely used tool to model relationships between objects belonging to various subject fields. It is possible to classify ontologies based on the degree of dependence on the task or application area, the model of ontological knowledge representation and expressiveness, as well as other criteria [22, 23].

To build ontology, first of all, we decided to define its subject field and scope. We answered some basic questions below:

1. What kind of fields does ontology include in? Answer: Noun.

2. For what we need ontology? Answer: To create a comparisonal ontological model according to noun in Kazakh and Turkish language.

3. What kind of questions does ontology answer? Answer: To define the types of affixes and to group according to the structure and meaning of noun.

4. Who does use and support ontology? Answer: The linguists and programmer. According to the answers, the ontological models that we need is following: O(X, R, I), X – the concepts that come into the structure of noun, R – relationships between these concepts, I – sets of the structure and relationships.

We used the ontology editor Protégé (http://protege.stanford.edu)to build the ontology. It is a free open source ontology editor and a framework for building knowledge bases. It was developed at Stanford University in collaboration with the University of Manchester. Concepts and relationships used in this ontological model are explained in Table 1.

Table 1

The concepts and relationships which was used in the ontological models of nouns for Kazakh and Turkish machine translation system

N	**Қазақша (Qazaqsha)**		**Türkçe**	**English**	**UNIFIED**
	Зат есім (Zat esim)		İsim	Noun	
	Құрамына қарай (Quramyna qaraj)			**Structure**	

1.1		Дара (Dara)		Simple	
1.2		Күрделі (Kuerdeli)		Complex	
1.2.1		Біріккен (Birikken)		Compound words	
1.2.2		Қосарланған (Qosarlanghan)		Reduplicates	
1.2.3		Тіркескен (Tirkesken)		Compound words	
1.2.4		Қысқарған (Qyskarghan)		Abbreviations	
	Тұлғасына қарай (Tulghasyna qaraj)		**Yapılarına göre**		
2.1		Негізгі (Negizgi)	Basit isim	Derıvations	DERIVNS
2.2		Туынды (Twyndy)	Türemiş ad	Derivatives	DERIV
2.3			Birleşik		
	Мағынасына қарай (Maghynasyna qaraj)		**Anlamlarına göre**	**According to the meaning**	
3.1		Жанды (Zhandy)	Canlı	Animate	ANIM
3.2		Жансыз (Zhansyz)	Cansız	Inanimate	INANIM
3.3		Жалпы (Zhalpy)	Cins isim	Common	COM
3.4		Жалқы (Zhalky)	Özel isim	Proper	PROP
			Oluşlarına göre		
		Деректі (Derekti)	Somut	Concrete	CON

		Дерексіз (Dereksiz)	Soyut	Abstract	ABS
	Санына қарай (Sanyna qaraj)		**Sayılarına göre**	**Number**	
5.1		Жекеше (Zhekeshe)	Tekil	Singular	SG
5.2		Көпше (Koepshe)	Çoğul	Plural	PL
5.3			Topluluk	Collectıve	
	ЗЕ жалғаулары (ZE zhalghaulary)		**Çekim ekleri**		
6.1	*Септік жалғауы (Septik zhalghauy)*		*Hal durum ekleri (İsim halleri)*	*Case ending*	*Cases*
6.1.1		Атау (Atau)	Yalın hali	Nominative case	NOM
6.1.2		Ілік (Ilik)		Genitive case	
6.1.3		Барыс (Barys)	Yönelme hali	Direction-dative case	DAT
6.1.4		Табыс (Tabys)	Belirtme (Yükleme) hali	Accusative case	ACC
6.1.5		Жатыс (Zhatys)	Bulunma hali	Locative case	LOC
6.1.6		Шығыс (Shyghys)	Ayrılma (çıkma) hali	Ablative case	ABL
6.1.7		Көмектес (Koemektes)		Instrumental case	
6.2	*Көптік жалғауы (Koeptik zhalghauy)*		Çoğul ekleri		Plural

6.2.1			Tekil	Singular	SG
6.2.2			Çoğul	Plural	PL
6.3	*Жіктік жалғауы (Zhiktik zhalghauy)*		Şahıs ekleri	Personal ending	Pers_end
		1 zhaq zhekeshe	Tekil I Şahıs	1 personal singular	PERS.1SG
		2 zhaq zhekeshe	Tekil II Şahıs	2 personal singular	PERS.2SG
		3 zhaq zhekeshe	Tekil III Şahıs	3 personal singular	PERS.3SG
		2 zhaq zhekeshe sypajy		2 personal singular formal	
		1 zhaq koepshe	Çoğul I Şahıs	1 personal plural	PERS.1PL
		2 zhaq koepshe	Çoğul II Şahıs	2 personal plural	PERS.2PL
		3 жақ көпше	Çoğul III Şahıs	3 personal plural	PERS.3PL
		2 zhaq koepshe sypajy		2 personal plural formal	
6.4	*Тәуелдік жалғауы (Taweldik zhalghauy)*		İyelik ekleri	Possesive ending	Poss_end
		1 zhaq zhekeshe	Tekil I İyelik	1 Possesive singular	POSS.1SG
		2 zhaq zhekeshe	Tekil II İyelik	2 Possesive singular	POSS.2SG
		3 zhaq zhekeshe	Tekil III İyelik	3 Possesive singular	POSS.3SG
		2 zhaq zhekeshe sypajy		2 Possesive singular formal	

		1 zhaq koepshe	Çoğul I İyelik	1 Possesive plural	POSS.1PL
		2 zhaq koepshe	Çoğul II İyelik	2 Possesive plural	POSS.2PL
		3 zhaq koepshe	Çoğul III İyelik	3 Possesive plural	POSS.3PL
		2 zhaq koepshe sypajy		2 Possesive plural formal	

Figure 1 illustrates the concepts and relationships which is used in the ontological models of nouns for Kazakh and Turkish machine translation system that used the notes

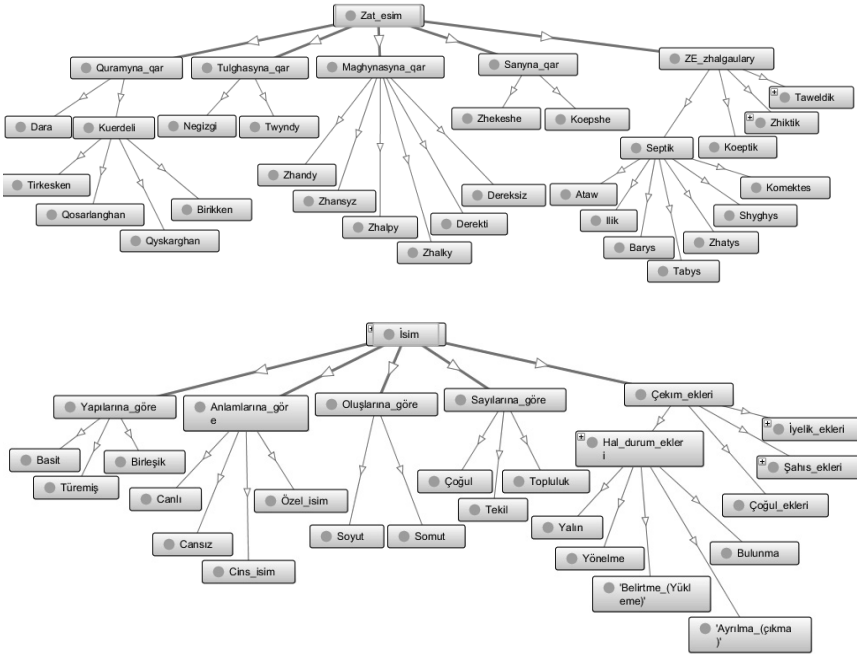


Fig. 1. The ontological models of nouns for Kazakh and Turkish machine translation system

In this way, the comparisonal ontological models of noun for machine translation system includes all the categories of morphological features, for instance, noun is divided as base and complex according to structure of noun in Kazakh language, whereas in Turkish language there is not such division, furthermore, a noun can be common, proper, concrete, abstract, animated, inanimate according to meaning in Kazakh language, while in Turkish language a noun can be common, proper, animated, inanimated. In both languages the divisions of affixation are similar, e.g, the forms of cases, number, possessives and conjugations. There are seven cases whereas in Turkish there are five. The similarities of both languages are illustrated as basic notes in Table 1.

Let's compare the similarities of cases and possessive forms of Kazakh and Turkish languages:

Table 2

Compare the common and possessive cases of Kazakh and Turkish languages

№	Cases		Common cases		Possessive cases	
	Kazakh	Turkish	Kazakh	Turkish	Kazakh	Turkish
	Atau septik	Yalın hali	Үй (Uej)	Ev	Үй**і** (Uej-i)	Ev**i**
	İlik septik		Үй**дің** (Uejding)			
	Barys septik	Yönelme hali	Үй**ге** (Uejge)	Ev**e**	Үй**іне** (Uejine)	Ev**ine**
	Tabys septik	Belirtme (Yükleme) hali	Үй**ді** (Uejdi)	Ev**i**	Үй**ін** (Uejin)	Ev**ini**
	Zhatys septik	Bulunma hali	Үй**де** (Uejde)	Ev**de**	Үй**інде** (Uejinde)	Ev**inde**
	Shyghys septik	Ayrılma (çıkma) hali	Үй**ден** (Uejden)	Ev**den**	Үй**інен** (Uejinen)	Ev**inden**
	Koemektes septik		Үй**мен** (Uejmen)		Үй**імен** (Uejimen)	

As we see, there are seven cases in Kazakh language, while five in Turkish language.

3. Conclusion

The creation of ontological models for computer processing of Kazakh and Turkish languages is the important step in comparative study of two Turkic languages. That is why the results of research and com-

parison of noun structure in similar Kazakh and Turkish languages will certainly have a bearing on the machine translation systems as well as the creation of processing systems of natural languages.

REFERENCES

[1] Kazakh grammar. (2002). Phonetics, word formation, morphology, syntax (in Kazakh). Astana.

[2] Кононов А.Н. Грамматика современного турецкого литературного языка. – М-Л: Издательство АН СССР, 1956. – 570 с.

[3] Yıldırım E., Tantuğ A.C., "The feasibility analysis of re-ranking for N-best lists on English-Turkish machine translation", IEEE International Symposium on Innovations in Intelligent Systems and Applications, Albena, Bulgaria, 2013.

[4] Yıldırım E., Tantuğ A.C., Evaluation of Domain Adaptation Approaches to Improve The Translation Quality, 6th International Conference on Computational Collective Intelligence Technologies and Applications (ICCCI 2014), Seoul, Korea, 2014.

[5] Tantuğ A.C., "Document Categorization with Modified Statistical Language Models for Agglutinative Languages", International Journal on Computational Intelligence Systems, vol. 5(3), 2010.

[6] M. Orhun, Tantuğ A. C., Adalı E., "Morphological Disambiguation Rules For Uyghur Language", IEEE International Conference on Software Engineering and Service Sciences (ICSESS), Beijing, China, 2010.

[7] Ilgen B., Adali E., Tantuğ A.C., A Comparative Study to Determine the Effective Window Size of Turkish Word Sense Disambiguation Systems, 28th International Symposium on Computer and Information Sciences, Paris, France, 2013.

[8] GulsenEryigit, Joakim Nivre, and Kemal Oflazer. Dependency parsing of Turkish. Computational Linguistics, 2008. – pp. 357–389.

[9] Umut Sulubacak and Gulsen Eryi˘git. Representation of morphosyntactic units and coordination structures in the Turkish dependency treebank. In Proceedings of SPMRL 2013 (4th Workshop on Statistical Parsing of Morphologically Rich Languages), Seattle, USA, October, 2013.

[10] Бекманова Г.Т., Махимов А.К. Графематический анализ казахского неструктурированного текста // Информатизация общества: труды 3-ой международной научно-практической конференции.– Астана, 2012. – С. 509–511.

[11] Бекманова Г.Т., Шарипбаев А. А. Формализация морфологических правил казахского языка с помощью семантической нейронной сети // Доклады Национальной академии наук Республики Казахстан. – 2009.– №4. – С. 11–16.

[12] Bekmanova G., Sharipbaev A. The synthesis of word forms of Turkic language using semantic neural networks // Modern problems of applied mathematics and information technologies: abstracts – Al Khorezmy, 2009. – P. 145.

[13] G.T.Bekmanova, A.A.Sharipbaeyv, S.K.Buribayeva Formalization of morphological rules of inflection in the Kazakh language // Вестник. Астана: Евразийский национальный университет им. Л.Н.Гумилева, 2012. – Специальный выпуск. – С. 18–26.

[14] A.Sharipbayev, G.Bekmanova, B.Yergesh, A.Buribayeva, and M.K.Karabalayeva. Intellectual morphological analyzer based on semantic networks. Processings of the OSTIS-2012, 2012, pp. 397–400.

[15] Sharipbaev A.A., Bekmanova G.T., Buribayeva A.K., Yergesh B.Z., Mukanova A.S., &Kaliyev A.K. (2012). Semantic neural network model of morphological rules of the agglutinative languages. In 6th International Conference on Soft Computing and Intelligent Systems and 13th International Symposium on Advanced Intelligence Systems, SCIS/ISIS, 1094–1099.

[16] Banu Yergesh, Assel Mukanova, Altynbek Sharipbay, Gulmira Bekmanova, and Bibigul Razakhova. Semantic Hyper-graph Based Representation of Nouns in the Kazakh Language. Computación y Sistemas Vol. 18, No. 3, 2014 pp. 627–635 ISSN 1405-5546 DOI: 10.13053/CyS-18-3-2041.

[17] Mukanova, A., Yergesh, B., Bekmanova, G., Razakhova, B., Sharipbay, A.Formal models of nouns in the Kazakh language. Leonardo Electronic Journal of Practices and Technologies.

[18] Sharipbayev A.A., Bekmanova G.T., Buribayeva A.K., Mukanova A.S. Semantic retrieval of information in the kazakh language in e-libraries. Journal of International Scientific Publications: Educational Alternatives, Volume 10, part 1, p.108–115. ISSN: 1313–2571, publishedat: http:// http://www.scientific-publications.net.

[19] ШарипбаевА.А., Разахова Б.Ш. Formalization of syntactic rules of the Kazakh language // Специальный выпуск. – Астана: ЕНУ им. Л.Н. Гумилева, 2012. – С. 42–50;

[20] Sharipbayev A.A., Razakhova B. Sh. Mathematical models of syntactical rules of Kazakh language subject to semantics of parts of the sentence // The 4th Congress of the Turkic World Mathematical Society (TWMS). – Baku, 2011. – p. 463;

[21] Шарипбаев А.А., Разахова Б.Ш. Определение множества предложений казахского языка с помощью контекстно свободной грамматикой. // Доклады Академии Наук Республики Казахстан. – Алматы, 2005. – №5. – С. 123–128.

[22] Gruber T.R. A Translation Approach to Portable Ontology Specifications / Gruber T.R.// Knowledge Acquisition,1993, P. 199–220.

[23] Gruber T.R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing / Gruber T.R. // International Journal Human-Computer Studies. – 1995, – Vol. 43 – P. 907–928.

THE ALGORITHM OF MACHINE TRANSLATION FROM UZBEK TO KARAKALPAK

Azizbek Kadirov

Nukus branch of the Tashkent University of Information Technologies
Nukus, Karakalpakstan, Uzbekistan

The problem of machine translation between Turkic languages is one of the important problems of linguistics and computer science. In this paper we examine an algorithm of machine translation from Uzbek to Karakalpak. An brief overview and analysis of grammar is given. Using the similarity of Uzbek and Karakalpak grammars, author presents online service of machine translation based on transformation and replacement algorithm. With appropriate upgrades, this algorithm can be used to create an automated system of machine translation between other Turkic languages.

В данной работе будет рассмотрена задача компьютерного перевода текстов с узбекского языка на каракалпакский и обратно.

Существует несколько форм организации перевода текстов при участии человека и компьютера:

- С предварительным редактированием – с целью приспособления текста для обработки компьютером, что позволяет исключить неоднозначности при переводе.
- С постредактированием – компьютер переводит текст, редактор вносит корректировки в результирующий текст.
- Интерактивный перевод – человек интерактивно принимает участие в переводе, разрешая сложные неоднозначные ситуации.
- Смешанные системы

Как правило, при компьютерном переводе специальная программа выполняет синтаксический анализ исходного текста, текст делится на предложения, предложения – на слова. Далее, определяется структура каждого предложения, после чего данная структура преобразуется в структуру предложения конечного языка. Но естественно, не всегда можно правильно определить, как построено предложение. Часто в тексте подразумевается некая мысль, которая должна быть понятна носителю языка, но которую программа просто не сможет уловить. Эти и другие обстоятельства препятствуют созданию полностью автоматизированной системы перевода текстов.

В Республике Каракалпакстан государственными языками являются каракалпакский и узбекский языки. Следовательно, большая часть документации ведется одновременно на двух языках – на каракалпакском, для внутреннего использования, и на узбекском, для отчетов и некоторых других документов.

Автором предложено использовать схожесть грамматики двух вышеуказанных языков для разработки системы компьютерного перевода текстов. Узбекский и каракалпакский языки, в числе прочих тюркских языков, относятся к типу агглютинативных языков, то есть формы слова образуются при помощи «приклеивания» к основе слова суффиксов и аффиксов. Сравнительный анализ показывает, что перевод текста с узбекского языка на каракалпакский можно алгоритмизировать.

Суть алгоритма состоит в следующем: исходный текст делится на слова. Далее, каждое слово исходного текста разбивается на морфемы, после производится поиск выделенных основ в базе данных и их перевод, после чего к переведенной основе в нужном порядке добавляют переведенные суффиксы и префиксы.

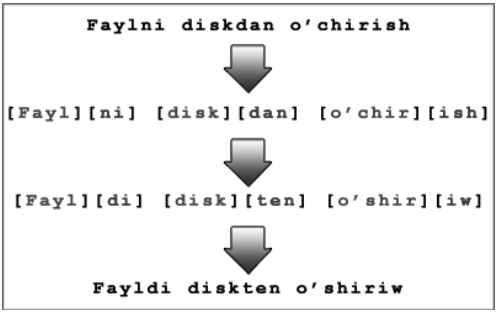


Рис 1. Алгоритм перевода (упрощённый вариант)

Каракалпакский язык по сравнению с узбекским имеет большее количество словообразовательных суффиксов. Например, узбекский суффикс *-lar*, служащий образованию множественного числа слова, имеет два аналога в каракалпакском языке – *-lar* и *-ler*. Использование того или иного суффикса зависит от мягкости предыдущего слога слова: после мягкого слога употребляется суффикс *-ler* (da'pterler – тетради, слог *ter* мягкий), после твердого

-*lar* (kitaplar – книги, слог *tap* твердый). Некоторые узбекские суффиксы имеют по 4 каракалпакских аналога, в этом случае выбор суффикса более сложен, и зависит произношения морфемы.

На базе данного алгоритма на сайте http://lugat.uz реализован онлайн перевод текстов с узбекского языка на каракалпакский. В базе данных программы хранится около 7.000 слов на узбекском и каракалпакском языках. После ввода исходного текста и команды на перевод алгоритм разбивает текст на слова, после чего каждое слово анализируется по приведенному выше алгоритму. Алгоритм переводит словосочетания «как есть», то есть слово в слово, в то время как многие словосочетания переводятся не на прямую. Для реализации функции перевода словосочетаний автор предлагает следующий алгоритм: создать базу словосочетаний на двух языках, затем, начиная с первого слова текста, сравнивать рядом стоящие пары слов со словосочетаниями из базы.

На данный момент ведется работа по расширению базы данных переводчика, а также по распознаванию различных словосочетаний. На входе алгоритм принимает тексты как на латинице, так и на кириллице, однако на выходе результат представляется в виде кириллицы. Для пополнения базы на сайте открыт модераторский раздел, где доверенные пользователи могут редактировать и добавлять новые слова.

Одной из проблем на данный момент является неоднородность базы. В связи с переходом каракалпакского языка на новую латиницу, необходимо заново проверить все слова в базе данных и соответственно обновить их написание. Часть данной работы была автоматизирована, но до сих пор много слов требуют ручной проверки. По этой причине некоторые слова на каракалпакский язык переводятся некорректно.

Приведенный довольно простой алгоритм, благодаря схожести множества тюркских языков, также может быть использован для автоматизированного перевода на казахский, таджикский и туркменский языки.

Использование электронного переводчика может оказать неоценимую помощь в самых различных сферах. Выходной текст, несмотря на неизбежную неточность перевода, можно использовать как черновик для дальнейшего редактирования профессиональным переводчиком. Во многих сферах, где не нужна абсолютная

точность перевода, и где важно передать основной смысл текста, электронный переводчик может стать удобным инструментом на рабочем столе пользователей.

ЛИТЕРАТУРА

Едемский М. Программы автоматического перевода //Мир образования. 1996. N 11–12. С. 54–55

Все о машинном переводе// ComputerBild. 2007. № 22

О внедрении каракалпакской письменности на основе латиницы (http://sovminrk.gov.uz/ru/pages/show/3323)

Кулагина О.С. О современном состоянии машинного перевода // Математические вопросы кибернетики, вып. 3, М.: Наука, 1991, стр. 5–50. Библиография из 140 названий.

Марчук Ю. Н. Проблемы машинного перевода. М.: Наука, 1983.

Интеллектуальные системы общего назначения http://www.intellsyst.ru/

http://www.promt.ru

http://www.socrat.ru

LEXICAL SELECTION RULES FOR KAZAKH-ENGLISH MACHINE TRANSLATION IN THE FREE/OPEN-SOURCE PLATFORM APERTIUM

Aidana Karibayeva

Information Systems Department, Al-Farabi Kazakh National University, 71 al-Farabi ave., Almaty, 050040, Kazakhstan
a.s.karibayeva@gmail.com

Kazakh language has great number of ambiguity words. The most of ambiguous word related to morphological ambiguity. The majority of homonyms pertain to morphological ambiguity. For example, “bas(бас)”, “kara(қара)”, “zhyz(жүз)”, “zher(жер)” and etc. All of these words related to two or more part of speech. The word “bas” can be as noun, verb, also word “zhyz” can be noun, numerical and verb. The word “zher” related to same part of speech like a word “bas”. Compared with other words this word has a lot of translation. This word has translation like this: “place”, “Earth”, “land”, “ground”, which are noun and “eat” as a verb in future time.

This paper describes process of building lexical selection rules for Kazakh-to-English machine translation system on free/open-source Apertium platform. Lexical selection rules are used for solving problems of ambiguity when ambiguity word has same part of speech. We will consider lexical selection rules for translating from Kazakh to English. Disambiguation is used to improve the quality of machine translation. This paper shows how to create lexical selection rules, what types of phrases and context are used to rules. Solving the task of ambiguity is a difficult task. Today, there are many tools of resolving it. One way of solving disambiguation is writing hand-written lexical selection rules, which we will consider at the paper.

In rule-based free/open-source platform Apertium disambiguation is solved by module of lexical selection. At module of lexical selection rules is written in XML-format. Lexical selection is used to determine the correct translation not the adequate sense. This difference differ it from word-sense disambiguation.

1. Introduction

Today ambiguity is a main problem of computer processing of language. So, each machine translation system must be solved this kind of tasks. Ambiguity appears when word of source language has a two or more translations in target language. In this paper we consider Kazakh

language as source language and English as a target language. These languages differ in syntax, morphology and they pertain to different type of language. Also, Kazakh as all Turkic language is agglutinative, whereas English is analytic.

Ambiguity can be lexical; morphological. Lexical ambiguity it means when ambiguity words have same part of speech, but by context it translated differently. Meanwhile, morphological ambiguity opposite to lexical. It means that in morphological ambiguity ambiguous word relate to different part of speech. In some case ambiguity calls polycemy. To solving problems of lexical selection is important to understand context which is translated to English.

Kazakh language has a great number of ambiguity words. We will consider Kazakh's words "bauyr (бауыр)" which is ambiguous word. By the nearby words we can distinguish meaning of this word. Firstly, it has a meaning like "brother", the second meaning is "liver". For example, "Менің бауырымның аты Дулат" and "Бауырым ауырып тұр". The first sentence is translated as "My brother's name is Dulat", the second sentence's translation is "Liver is hurts".

Not only noun has ambiguity, pronoun also can be ambiguous. The pronoun "Ол" has tree translations. It can be "he", "she" and "it". When this word appears in context which is meant female, this translated as "she". In Kazakh this word ambiguous, but in English it is unambiguous.

By considering features of translation by context, we are developing rules which are solving the task of ambiguity from Kazakh to English based on the Apertium free/open-source machine translation platform (Forcada et. al., 2011). For solving polycemy of Kazakh–English language pair we need to build bilingual dictionary and write couple of lexical selection rules.

This paper contains 4 sections: Section 2 describes Apertium platform and its structure, Section 3 describes Kazakh–English lexical selection and Section 4 gives results of system.

2 Apertium platform and its modules

Apertium is a free/open source machine translation system. Apertium is free software which is published by developers according to GNU GPL conditions (Apertium).

At the first time Apertium was developed for translation between similar languages. However this system has been expanded to translate texts between dissimilar language pairs, such as English – Kazakh (vice versa) language pairs. For developing we need to create the linguistic data (dictionaries, rules), which are written in XML formats. By using dictionaries we find words which have two or more translations. So, this machine translation system uses finite state transducers for all of its lexical transformations, and hidden Markov models for part-of-speech tagging or word category disambiguation.

This machine translation system has own modules for implementation of transfer.

Apertium platform consists following modules (Sundetova et.al., 2014):

- Deformatter
- Morphological analyser
- Part of speech tagger
- Lexical transfer
- Lexical selection
- Structural transfer
- Morphological generator
- Post-generator
- Re-formater

So, we consider how to work these modules. First module is **deformater**. This module divides the source text to formatting tags. These tags called as "superblanks" which insert the place between words.

Second module is **morphological analyser**. Morphological analyser constitute to each lexical unit one or more lexical forms. These form consist of lemma, lexical category or part of speech. Morphological analysis is generated by compiling a morphological dictionary of source language. Lexical units containing more than one word (multiword lexical units) are analyzed as a single lexical unit. Morphological analyser uses a finite state transducer based on two-level rules (in the case of Kazakh, apertium-kaz.kaz.lexc, apertium-kaz.kaz.twol). This module therefore separates lexemes and processes morphological analysis, and then returns possible lexical forms. Below we show the morphological analysis of ambiguous word.

```
^жүз/жүз<num>/жүз<n><nom>/жүз<n><attr>/
жүз<num><subst><nom>/жүз<v><iv><imp><p2><sg>/
жүз<n><nom>+e<cop><aor><p3><pl>/жүз<num><subst>
<nom>+e<cop><aor><p3><sg>/жүз<num><subst><nom>+
e<cop><aor><p3><sg>$^./.<sent>$
```

As you see this word has 9 morphological interpretations. The frequent interpretation is as a numerical. There four analysis with numerical: `жүз<num>`, `жүз<num><subst><nom>/`, `жүз<num><subst><nom>+e<cop><aor><p3><pl>/`,`жүз<num><subst><nom>+e<cop><aor><p3><sg>`. Here <num> – numerical, <attr>- attributive, <nom>- nominative, <cop>-copula, <p3>- third person and etc (List of symbols: wiki.apertium.org/wiki/List_of_symbols). We receive three analysis with noun, in addition the noun can be nominative, attributive and nominative with copula. Although, we have only one analysis with verb. This analysis like this: `жүз<v><iv><imp><p2><sg>`, where "zhyz" is verb, intransitive, imperative, second peson and singular. By context we distinguish the corresponding analysis using the part of speech tagger, which we consider below.

Third module is **part of speech tagger** which is based on hidden Markov model(HMM). Final result of part of speech we receive after applying constraint grammar rules. In Kazakh directory this file of rules called "apertium.kaz.kaz.rlx". Here we solve the morphological ambiguity. This type of polysemy appear when the word of source language can be relate two or more part of speech. For example, word "zhyz(жүз)" can be relate to tree part of speech, namely it can be translated as a noun, numeral and verb. If we consider the sentence or phrase "zhyz tenge" it translated as "hundred tenge". So, in this context this word's part of speech is numeral. After applying these rules we receive just one morphological analysis. Here we show the rule for this construction:

```
SELECT Num IF
    ((0 Num) OR (-1))
    (1 N)
```

This rule shows that we choose "zhuz" as a numeral if this word come with numeral or noun before of source word.

Forth module is **lexical transfer**. This module works with bilingual dictionary (apertium-eng-kaz.eng-kaz.dix) (Сундетова, Кәрібаева, 2013), from this dictionary lexical transfer module reads lexical form of source language and retrieve corresponding lexical form of source language. The lexical forms of target language can be one or more than two.

Fifth module is lexical selection which we consider in this paper. This module uses the lexical selection rules. The rules is written by hand in file apertium-eng-kaz.kaz-eng.lrx by determining nearest word or context. So, lexical ambiguity is solved here. Lexical selection is the focus of this paper, so we described in detail in the next section.

Sixth module is **structural transfer.** This module uses to transform source language sentence or phrase to target language by using transfer rules. This module covers syntactic processing. To processing it uses transfer rules, which transform lexical forms sequences to another sequence of target language. Structural transfer works in tree step. First of all is "chunker" level, which divide source sentence to chunks. At the second level, namely in "interchunk" it did rearrangement of phrases. For example: "Мен/SN бақшада/SN-LOC ойнаймын/SV" translated to English as "I/SN play/SV in garden/SN-LOC". Here "SN" means noun phrase, "SV" is verb phrase. As you see Kazakh language has "SOV" type, whereas English is "SVO". So, "interchunk" level did arrangement from "SOV" to "SVO". At final level it does some clean-up by deleting unnecessary tags.

Seventh module is **morphological generator.** It generates a corresponding sequence of target language surface forms. The morphological generator executes a finite-state transducer generated by compiling a morphological dictionary for the target language.

The penultimate module is **post-generator**. It takes care of some minor orthographical operations in the target language.

Last module is **reformatter.** It places format tags back into the text so that its format is preserved.

3. Lexical selection from Kazakh into English languages

The lexical selection module is the one of main module in receiving correct translation. The lexical selection module in Apertium does disambiguation, namely solving task of lexical ambiguity.

The operations, which is used in writing rule show below in the Table1.

Table 1

Operations of lexical selection rules

Operations	Meaning
<rule>	Start of rule
<SELECT>	Operation of choosing
<tags>	Determine corresponding tag to word
<match>	choose
<lemma>	Lexical form
</rule>	End of rule

English-Kazakh and Kazakh-English language pairs use same linguistic data of dictionaries. These dictionaries are monolingual dictionary of English, lexical dictionary of Kazakh and bilingual dictionary of both languages. They differ by number of words there. The monolingual Kazakh dictionary consist about 20000 words, the monolingual dictionary of English 36876 and bilingual consist 13751 words (current version: 50582)

By adding words to dictionary, it increased number of ambiguity. Kazakh language is a rich language with ambiguity. We present some words which have several translations from bilingual dictionary. There are some ambiguous words with its translation (Table 2):

Table 2

Example of ambiguous words with its translation

Kazakh words	Translations	POS of translation	Example with context	Translation by using context
бет	page, face, surface	noun	адамның беті	face of people
			кітаптың беті	book page
			судың беті	water surface
ол	he, she	pronoun	Ол қыз	She is girl
			Ол бала	He is boy

үй	home, house	noun	үйге бару	go to home
			үйде тұру	live in house
оқу	read, study	verb	университетте оқу	study at university
			кітапты оқу	read a book
жастық	pillow, adolescence, youth	noun	жастыққа жату	lie on the pillow
			жастықтықты еске алу	remember the adolescence
ара	bee, saw	noun	арамен кесу	To pick the tree by saw
			ара шағып алды	bee bite

3.1. The Kazakh-English lexical selection rules

Kazakh language has no gender. In sentence, which has personal pronoun “ol” we must do lexical selection. So, we solve this by writing the rules in lexical selection file of Kazakh into English machine translation system. So, this rule will be like this:

```
<rule>
   <match    lemma="Ол"    tags="prn.pers.p3.sg.
nom"><select    lemma="she"    tags="prn.subj.
p3.f.*"/></match>
   <match lemma="қыз" tags="n.*"/>
</rule>
```

This rule starts with matching the lemma and its tags. The tags illustrate morphological analysis. “prn” is pronoun, “pers” is personal, “p3” is third person, “sg” is singular and “nom” is nominative. After determining tags, we choose lemma, which must be corresponded and write its tags. If we don’t write rule for this case, the system translated all sentence with “ol” as he. So, we generate “he” by default.

The next example of rule illustrated the translation with verb. The word “oku(оқу)” has two corresponding translation. It translated as “read” and “study”. When this word come with “university” or “institute”, it translated as “study”. Meanwhile, by default it has a transla-

tion “read”. Below we illustrate the rule for the first case, which we discuss:

```
<rule>
  <or>
   <match lemma=»институт» tags=»n.loc»/>
   <match lemma=»университет» tags=»n.loc»/>
  </or>
   <match     lemma=»оқы»     tags=»v.*»><select
lemma=»study» tags=»vblex.*»/></match>
</rule>
```

In the case when context or nearest word connected with university or institute, it translated as a “study”.

Rules for this phrase are assigned to verb case. After analyzing lexical selection we see that context has a great importance in generating rules. In this level of lexical selection rules are written 15 rules. There are some rules from lexical selection file:

```
<rule>
   <match     lemma=»Ол»     tags=»prn.pers.p3.sg.
nom»><select     lemma=»she»     tags=»prn.subj.
p3.f.*»/></match>
   <match lemma=»қыз» tags=»n.*»/>
 </rule>

 <rule>
   <match     lemma=»Ол»     tags=»prn.pers.p3.sg.
nom»><select     lemma=»she»     tags=»prn.subj.
p3.f.*»/></match>
   <match     lemma=»әдемі»     tags=»adj»><select
lemma=»beautiful» tags=»adj»/></match>
   <match lemma=»қыз» tags=»n.*»/>
 </rule>

 <rule>
   <match lemma=»кітап» tags=»n.*»/>
   <match     lemma=»бет»     tags=»n.*.*»><select
lemma=»page» tags=»n.*»/></match>
 </rule>
```

```
  <rule>
   <match     lemma=»әдемі»     tags=»adj»><select
lemma=»beautiful» tags=»adj»/></match>
 </rule>

 <rule>
   <match    lemma=»оның»    tags=»prn.pers.p3.sg.
gen»><select  lemma=»his»  tags=»det.pos.sp»/></
match>
 </rule>

 <rule>
   <match      lemma=»үй»      tags=»n.*»><select
lemma=»home» tags=»n.*»/></match>
   <match lemma=»*» tags=»v.*»/>
 </rule>

 <rule>
   <match      lemma=»үй»      tags=»n.*»><select
lemma=»house» tags=»n.*»/></match>
 </rule>

 <rule>
   <match    lemma=»арқылы»    tags=»post»><select
lemma=»through» tags=»pr»/></match>
 </rule>

 <rule>
  <or>
   <match lemma=»институт» tags=»n.loc»/>
   <match lemma=»университет» tags=»n.loc»/>
  </or>
   <match      lemma=»оқы»      tags=»v.*»><select
lemma=»study» tags=»vblex.*»/></match>
 </rule>
```

4. Results

The current version of the system (revision №60582) by hand-rules can decide ambiguity noun, pronoun and verb – phrases. We plan to extend the number of rules to improve translation quality.

Here we show some results of translation, see Fig. 1,Fig. 2,and Fig. 3.

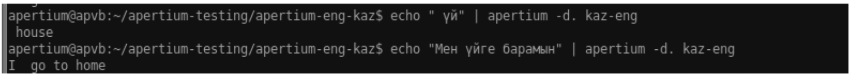


Fig. 1. Result of translating ambiguous noun.

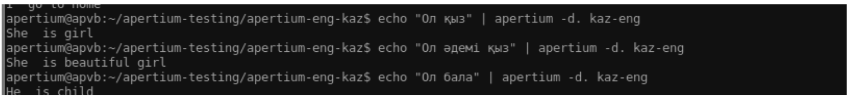


Fig. 2. Result of translating ambiguous pronoun.

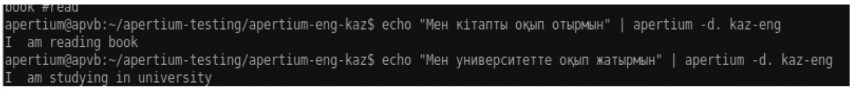


Fig. 3. Result of translating ambiguous verb

5. Conclusion

We have described Kazakh–English machine translation system on Apertium platform and process of solving disambiguation. Many features in translating from Kazakh to English as selection cases of noun, verb, pronoun and etc. were solved. However, hand-written lexical selection rules do not cover all situations with ambiguity, because before writing the rules, we must find ambiguity words in context, which require few times. So, we must create a new tool, which generate this kind of rules automatically. In the future this system will be considered automatically generation of lexical selection rules.

This research are conducted under grant funding 0749 / GF4 of the Ministry of Education and Science of the Republic of Kazakhstan.

REFERENCES

Forcada, M.L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J.A.

Sánchez-Martínez, F., Ramírez-Sánchez, G., Tyers, F.M. 2011. *"Apertium: a free/open-source platform for rule-based machine translation". Machine Translation 25(2)127–144.*

Apertium (2015). Retrieved from http://en.wikipedia.org/wiki/Apertium.

Sundetova A, Karibayeva A., Tukeyev U.A. *STRUCTURAL TRANSFER RULES FOR KAZAKH-TO-ENGLISH MACHINE TRANSLATION IN THE FREE/OPEN-SOURCE PLATFORM APERTIUM.* The International Conference on Turkic language processing "TURKLANG'14", Istanbul Technical University, 6–7 November 2014, – 91–96 p.

List of symbols (2015). Retrieved from http://wiki.apertium.org/wiki/List_of_symbols.

Сундетова А.М., Кәрібаева А.С., *Апертиум платформасындағы Ағылшын–Қазақ машиналық аудармашы үшін екітілді сөздікті құру.* Материалы международной научно-практической конференции «Применение информационно-коммуникационных технологий в образовании и науке», посвященной 50-летию Департамента информационно-коммуникационных технологий и 40-летию кафедры «Информационные системы» КазНУ им. аль-Фараби. 22 ноября 2013 г. – Алматы: Қазақ Университеті, 2013. – С. 53–57.

REALISATION OF STATISTICAL MACHINE TRANSLATION BASED ON A PARALLEL TATAR-RUSSIAN CORPUS OF LEGAL TEXTS

Aliya Mirzagitova

St. Petersburg State University, St. Petersburg, Russia
amirzagitova@gmail.com

This paper considers the problems and details of realisation of a basic statistical machine translation system for Tatar-Russian language pair based on a parallel corpus of a restricted domain. In spite of growing interest in automatic processing of Tatar language and corpus linguistics in general, there are still no available Tatar-Russian parallel corpora which could enable independent researchers to experiment on large amounts of texts. The chief application of such resources lies in the area of machine translation. Though there are several publications on successful development of rule-based machine translators for Tatar in pair with its cognates, we still chose statistical approach due to assumption that it could help to bypass the systematic differences between agglutinative and inflectional morphologies of Tatar and Russian respectively.

Current linguistic situation in the Republic of Tatarstan (particularly, Tatar being one of the two state languages) allowed us to collect a fair amount of official translations of legal texts, which are precise by definition and characterized by high quality. We aligned collected sentences and used this parallel corpus in order to implement the automatic translation system by means of a free and open-source instrument Moses that comprises most of the state of the art techniques, including phrase-based statistical machine translation. Moreover, we suggest an approach for automated extraction of possible candidates into a bilingual Tatar-Russian dictionary of words and terms concerning the restricted domain of special texts. The proposed method uses the information from phrase tables and could therefore potentially enrich the existing linguistic tools and resources in a straightforward manner.

Введение

За свою 65-летнюю историю машинный перевод переживал взлёты и падения. Сегодня можно утверждать, что он вновь испытывает подъём – регулярно публикуются новые результаты исследователей из разных стран, создаются и улучшаются открытые и коммерческие системы для различных пар языков. Несмотря на то,

что исторически основное внимание до недавнего времени уделялось созданию моделей для европейских языков, современное состояние машинного перевода позволяет разрабатывать подобные продукты для самых различных языковых пар, в том числе и для языков из разных семей и групп. Особое значение имеет создание открытых инструментов для малоресурсных языков.

В последние годы в Татарстане активно ведутся разработки в области компьютерной лингвистики: создаются татарские корпуса, предлагаются новые методы лингвистической разметки. В сфере машинного перевода большие успехи были достигнуты для татарского в паре с другими тюркскими языками (Сулейманов и др., 2014). С другой стороны, несмотря на очевидную сложность, заключающуюся в отличиях между агглютинативными и флективными языками, востребовано татарско-русское и русско-татарское направления перевода. Это объясняется в первую очередь языковой ситуацией в республике: татарский действует как второй государственный язык (законы издаются на двух языках), его преподают школах (в том числе, печатаются учебники и их переводы), функционируют двуязычные СМИ. Перечисленные факторы предоставляют возможности для создания параллельных корпусов и разработки систем статистического машинного перевода, в которых можно было бы преодолеть системные различия между татарским и русским.

Насколько нам известно, в данный момент отсутствуют какие-либо публикации о разработке статистических машинных переводчиков для татарско-русской языковой пары. Более того, несмотря на возросший интерес, на данный момент в открытом доступе нет параллельных корпусов, которые могли бы дать исследователям возможность проводить различные эксперименты, в первую очередь, по автоматическому переводу.

По этой причине мы поставили перед собой задачи разработки татарско-русского параллельного корпуса специальных текстов, а также создания базовой системы машинного перевода и проверки гипотезы о возможности автоматического извлечения двуязычного татарско-русского лексикона для последующего решения задач машинного перевода и автоматического составления различных типов словарей.

2. Методы статистического машинного перевода

В компьютерной лингвистике сформировались три подхода к автоматическому переводу: основанный на правилах, статистический и гибридный. Несмотря на высокую точность, перевод на правилах для пары агглютинативного и флективного языков представляет собой трудоёмкую задачу. В нашем исследовании мы обратились ко второму подходу.

Статистический подход возник в начале 1990-ых годов, и на сегодня это наиболее активно развивающаяся архитектура машинного перевода. Он позволяет извлекать знания о языке, основываясь только на больших массивах данных. К его плюсам относится относительно малозатратное построение переводчиков при наличии параллельных корпусов. Минусы заключаются в ограниченной доступности корпусов достаточного объёма, способных покрыть всю агглютинативную морфологию. При отсутствии наложенных сверху правил (гибридный подход) только увеличение репрезентативности корпусов может скомпенсировать недостаточное отражение в модели машинного перевода морфологии и синтаксиса конкретного языка.

В целом, методы статистического машинного перевода объединяет то, что применяется статистический анализ параллельных корпусов. Любая система статистического машинного перевода состоит из трёх главных компонентов: языковой модели, модели перевода и декодера. В основе языковой модели лежит распределение вероятностей различных последовательностей слов. Модель перевода оценивает вероятность появления исходного предложения при условии данного целевого предложения.

В ранних алгоритмах статистического перевода за единицу текста принималось слово. Иначе говоря, одно слово на исходном языке переводилось одним словом на целевом языке. Подобный подход имеет очевидный недостаток: исходное слово может переводиться несколькими словами и наоборот, несколько исходных слов могут переводиться одним словом на целевом языке. По этой причине в модели перевода по фразовым таблицам за единицу берутся короткие непрерывные последовательности слов. Процедура перевода при этом состоит из нескольких шагов: сегментация

предложения на фразы; перевод каждой фразы; переупорядочение переведённых фраз.

Собственно переводом занимается декодер, находящий перевод, одновременно максимизирующий произведение вероятностей из модели перевода и языковой модели.

3. Создание татарско-русского параллельного корпуса юридических текстов

На данный момент успешно создаются инструменты для автоматической обработки татарского языка и собираются необходимые для этого ресурсы в виде текстовых корпусов. Следует упомянуть национальный корпус татарского языка «Туган тел» (web-corpora.net/TatarCorpus) в 20 млн слов, письменный корпус татарского языка (corpus.tatfolk.ru/) в 116 млн слов. Тем не менее, они одноязычны и доступны только для онлайн пользования, что ограничивает их применение в рамках поставленной задачи. По этой причине было принято решение собрать параллельный корпус по доступным источникам.

В Республике Татарстан законы и иные нормативные правовые акты публикуются на русском и татарском языках. Из этого можно было бы заключить, что параллельный корпус из текстов официальных документов, размещённых на русской и татарской версиях сайта правительства РТ (gossov.tatarstan.ru), будет содержать надёжные переводные эквиваленты, подтверждённые юридически.

Особенность такого источника текстов заключается в том, что они принадлежат узкой специализированной области знаний и тем самым содержат в основном только ограниченную в использовании лексику и специфические синтаксические конструкции. Из-за этого многие бытовые слова оказываются вне словаря, что не позволяет переводить тексты не из данной предметной области. Однако по опыту признанных специалистов в области машинного перевода, «подобные тексты могут использоваться для создания или уточнения нормативных словарей соответствующих областей знаний» (Беляева, 2004).

Таким образом, использование текстов подобной узкой тематики обосновывается, во-первых, высоким качеством и надёжностью перевода; во-вторых, наличием только этих данных в открытом доступе. В-третьих, пользование ими не ограничено авторскими правами.

В качестве материалов для корпуса были выбраны: Конституция Республики Татарстан, Декларация о Государственном суверенитете Татарской Советской Социалистической Республики, Регламент Государственного Совета Республики Татарстан, несколько официальных изданий «Ведомостей Государственного Совета Татарстана». Последние представляют собой собрания нормативных документов, изданных за определённый период времени, обычно один месяц.

Документы имеют типовую структуру и дублирующиеся составные части, что, очевидно, влияет на частотные характеристики n-грамм. Например, в любом документе можно выделить: тип (закон, распоряжение и т.д.); название – чаще шаблонное («Об обращении Законодательного Собрания…»); дату принятия; текст, который может быть разделён на пронумерованные статьи и пункты статей; подпись; место и дату. Общий объём корпуса составляет приблизительно 315 тыс. слов для каждого языка.

Предподготовка исходных данных включала в себя сегментацию на уровне предложений с помощью регулярных выражений. Для выравнивания корпуса использовался открытый, свободный для скачивания инструмент LF Aligner (sourceforge.net/projects/aligner). В основе лежит алгоритм, строящий псевдо-словарь при помощи информации о длине предложений.

Информация о количестве выровненных предложений представлена в таблице 1.

	Татарский язык	Русский язык
Предложения	15980	
Токены	312208	312616

4. Экспериментальная система статистического машинного перевода

Для создания базовой системы татарско-русского машинного перевода был использован открытый инструмент Moses (statmt.org/moses) (Koehn, 2007), в котором реализованы основные современные методы статистического машинного перевода. Moses включает в себя компонент для обучения, декодер, а также множество вспомогательных инструментов для подготовки корпусов.

Для обучения нашей системы было взято следующее количество предложений (таблица 2).

Таблица 2

Параметры обучающей выборки

	Татарский язык	Русский язык
Предложения	13168	
Токены	243513	244888

Перед процедурой обучения была проведена обработка корпуса: токенизация, приведение к единому регистру, удаление слишком длинных предложений.

Затем в параллельных предложениях проводился поиск соответствий между токенами, т.е. производилось выравнивание на уровне слов при помощи GIZA++ (Och, 2003). Проводится симметризация, т.е. процедура применяется в обоих направлениях, эвристикой добавляются новые соответствия в пространстве между пересечением и объединением двух таблиц (Tian, 2014). Параллельно создаётся файл с прямыми переводами для отдельных слов, вероятности которых приближаются методом максимального правдоподобия. Фрагмент такого файла для слова *ант (присяга)*, где отражены найденные в текстах переводы на русский и их вероятности (NULL обозначает пустой токен): *ант принесения 0.1111111; ант присяги 0.9000000; ант приносит 0.5000000; ант присяге 0.5000000; ант присягу 0.8000000; ант NULL 0.0001326.*

В дальнейшем результаты выравнивания применяются для построения фразовой таблицы, элементами которой являются фраза на татарском языке, её перевод на русский язык и вероятности перевода: *аларның гамәлгә ашырылуын тәэмин итәргә ||| обеспечить их реализацию ||| 0.777778 0.111111 0.111111 0.777778 0.111111 0.111111*

Языковую модель машинного перевода строят на части параллельного корпуса для целевого языка. По умолчанию для построения модели применяется уже встроенный в систему Moses инструмент KenLM.

Заключительный шаг разработки экспериментальной системы машинного перевода заключается в уточнении полученных вероятностей перевода для нахождения наиболее приемлемого пере-

вода. На этом шаге используется методика MERT (Minimum Error Rate Training), суть которой заключается в стремлении увеличить значение BLEU на корпусе, используемом для настройки и для которого требуются параллельные тексты, не использованные ранее при обучении. Для этого были случайно отобраны 990 предложений (около 23 тыс. токенов). Сначала эта выборка переводится, затем результат сравнивается с эталонным переводом и вычисляется значение оценки BLEU. Соответственно величины, приводящие к лучшей оценке, записываются в файл конфигурации. После этого проводится второй шаг с новыми значениями – повторно переводится выборка, сравнивается с эталоном, вычисляется значение BLEU. Цикл повторяется до тех пор, пока оцениваемые значения не стабилизируются.

4.1. Оценка результатов

Изначально для тестирования были отобраны два вида текстов: сборник законов с характерным для экспериментального корпуса стилем и типовым словарём (21 тыс. токенов); и несколько отличающийся по лексике и структуре Бюджетный кодекс РТ (25 тыс. токенов). Решение оценить работу на отличающемся материале было принято из-за высокой степени схожести сборников и, как следствие, завышенных результатов оценивающей метрики.

Самая распространённая оценка машинного перевода – это BLEU, в основе которой лежит предположение, что чем ближе автоматический перевод к человеческому, тем он лучше (Papineni, 2002). Несмотря на грубость критерия, эта метрика получила наиболее широкое распространение и служит постоянным ориентиром при описании создаваемых систем. Для первого текста был получен результат 18; для второго, более общего, текста – 10,68.

Пример базового перевода двух смежных предложений приведён в таблице 3.

Таблица 3

Пример машинного перевода

Оригинал	дүртенче чакырылыш татарстан республикасы дәүләт советының илленче утырышы каравына кертелә торган мәсьәләләр турында татарстан республикасы дәүләт советы президиумы карар бирә :

Перевод	о вопросах, вносимых на рассмотрение **пятидесятого** заседания государственного совета республики татарстан четвертого созыва президиум государственного совета республики татарстан постановляет :
Машинный перевод	комитету государственного совета республики татарстан четвертого созыва илленче заседания о вопросах, вносимых на рассмотрение президиум государственного совета республики татарстан постановляет :

Как видно, помимо неправильного порядка слов в нешаблонных фразах (ср. с *президиум ... постановляет*), базовая система не может перевести слово вне словаря (*илленче – пятидесятый*). На данный момент нам кажется, что без использования лингвистических инструментов эти проблемы можно решить лишь за счёт увеличения обучающего корпуса с большим покрытием как бытовых слов, так и юридических терминов.

5. Автоматическое извлечение двуязычного татарско-русского лексикона

В ходе экспериментов было выдвинуто предположение, что фразовая таблица, создаваемая на этапе обучения системы машинного перевода, может послужить источником переводных эквивалентов. Несмотря на то, такая разновидность словаря вряд ли найдёт широкое применение среди пользователей, подобный ресурс способствовал бы, в первую очередь, развитию области автоматической обработки татарского языка на фоне дефицита электронных словарей. Наличие двуязычного лексикона оказалось бы полезным в разработке систем машинного перевода, основанного на правилах, где наблюдается большой недостаток татарско-русских и русско-татарских словарей достаточного объёма для покрытия разных предметных областей.

Во-вторых, в случае разработки новых или улучшения существующих лингвистических инструментов для татарского языка, было бы удобно иметь возможность получить список несловарных лексем с их эквивалентом на русском языке (например, именованные сущности). Тем самым, создание такого списка способство-

вало бы доработке автоматического морфологического анализа, в частности, усовершенствованию алгоритмов предсказания.

Специалисты признают, что автоматически созданные двуязычные лексиконы способны отразить те значения, которые часто неочевидны даже для лингвистов и носителей языка. Содержание лексиконов можно регулярно обновлять, чтобы добавлять новые параллельные тексты и тем самым находить новые появившиеся слова и выражения (Antonova Misyurev, 2014).

Сформированная нами на предыдущем этапе исследования фразовая таблица изначально содержала около 200 тыс. записей. Очевидно, среди них большую часть составлял шум, который необходимо было уменьшить. Из-за отсутствия в открытом доступе морфологических и синтаксических анализаторов для татарского языка, было принято решение установить пороговое значение величины вероятностей (Koehn et al., 2003). Для того чтобы ещё сильнее сократить разброс, в лексиконе были оставлены только униграммы, т.е. только одно слово и все его возможные переводы с вероятностями выше порогового значения. Это было сделано потому, что без синтаксического парсера невозможно достоверно классифицировать выделенные n-граммы как грамматически правильные и неправильные (Antonova Misyurev, 2014): например, переводная пара *иминлек һәм – безопасности и* не представляет интереса. После этих процедур в словаре осталось 2556 слов.

Затем мы обратились к морфологическому анализатору из открытой платформы машинного перевода на правилах Apertium (apertium.org), основанному на конечных преобразователях и словаре в 12 тыс. слов. Была проведена лемматизация оставшихся униграмм, из них 935 словоформ остались неразобранными (36,5%).

Далее лемматизированные униграммы были проверены на вхождение в двуязычный словарь из Apertium, в котором на тот момент имелось 6000 переводных пар. Новыми были 662 извлечённых слова с переводами, тем самым, за счёт этих переводных пар можно было расширить существующий словарь.

В качестве примера можно привести некоторые из выделенных соответствий, которых до этого не было в готовом словаре: *айлык – ежемесячный, алмаш – представительство, архангел – архангельский, баз – рынок, барс – барс, беркетмә – протокол, вазыйфа – должность, вәзгыять – ситуация, журналист – жур-*

налист, камилләштерү – совершенствование, кичекмәстән – незамедлительно, кодекс – кодекс, компетенция – компетенция, мөэмин – мукмина, партнерлык – партнерство, политехник – политехнический, президентлык – президентский, хосусыйлаштыру – приватизация, цех – цех, читләш – уклонение, чуваш – чувашский, ялан – елань, һәлак – погибший.

6. Заключение

Главной целью проведённых экспериментов было построение татарско-русского статистического машинного переводчика для узкоспециальной предметной области (законодательных текстов) на базе свободного инструмента Moses. Для этого был собран новый татарско-русский параллельный корпус объёмом в приблизительно 300000 токенов для каждого языка, проведена статистическая обработка текстов корпуса. В результате автоматической оценки качества экспериментальной системы машинного перевода при помощи метрики BLEU были получены достаточно высокие значения, однако следует иметь в виду, что представленные тексты относятся к узкой предметной области. Анализ ошибок показал, что в перспективе необходимо расширять корпус текстов и рассмотреть способы сокращения фразовой таблицы (например, применение лемматизации, стемминга).

В ходе экспериментов был извлечён татарско-русский словарь униграмм на базе фразовой таблицы из построенной системы машинного перевода. Были рассмотрены возможности применения подобного словаря для улучшения существующих лингвистических инструментов для автоматической обработки татарского языка.

Помимо увеличения корпуса, перспективы исследования связаны с проведением более точной оценки, а также количественным и качественным сравнением с недавно появившейся русско-татарской парой в сервисе Яндекс.Переводчик (translate.yandex.ru).

ЛИТЕРАТУРА

Беляева Л.Н. (2004). *Лексикографический потенциал параллельного корпуса текстов.* Труды международной конференции «Корпусная лингвистика – 2004». СПб.

Сулейманов Д.Ш., Гатиатуллин А.Р., Гильмуллин Р.А., Аюпов М.М. (2014). *Система машинного перевода для тюркских языков.* Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2014. Казань: Фэн.

Antonova A., Misyurev A. (2014). *Improving the precision of automatically constructed human-oriented translation dictionaries.* EACL 2014.

Koehn P., Och F.J., Marcu D. (2003). *Statistical Phrase-based Translation.* Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Stroudsburg: ACL.

Koehn, P. (2007). *Moses: Open Source Toolkit for Statistical Machine Translation.* Annual Meeting of the Association for Computational Linguistics (ACL). Prague.

Och, F. J., Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29 (1).

Papineni K. et al. (2002). *BLEU: a method for automatic evaluation of machine translation.* Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics.

Tian L., Derek F.W., Lidia S.C., Oliveira F. (2014). *A Relationship: Word Alignment, Phrase Table, and Translation Quality.* The Scientific World Journal.

THE HISTORY OF TRANSLATION IN YAKUTIA : ACHIEVEMENTS AND PROBLEMS

Alina Nakhodkina

M.K. Ammosov North-Eastern Federal University

The article reviews the history of translation, in particular Russian-Yakut translation, in Yakutia – a northeastern region of Russia – since 17th century till nowadays. The author pays attention to the status of the Yakut (Sakha) language as "an international one… in all north-east of Siberia" including the 19th century. The factors that stimulated much the process of translation in Yakutia are as follows: the missionary activity of the Russian Orthodox Church; political exiles; the Great October Socialist Revolution and Soviet policy of the liquidation of illiteracy; as well as Perestroika (Reconstruction) and post-reconstruction period. Lexicographic works are one of the typical signs of the initial stage of intercultural interaction, according to the author, while translation from Yakut into Russian and foreign languages; interpretation, especially simultaneous translation from and to Yakut; and foundation of the professional union of translators characterize the advanced level of translation in Yakutia. There are analyzed the achievements and challenges of the modern period since 1990s connected with the languages of the Yakut and indigenous peoples of the North as a result of the language policy in the Russian Federation.

Известно, что изучение и осмысление литературы, культуры и искусства любого народа независимо от его места в истории цивилизации, проходят очень сложный и длительный путь. И то, что история якутской литературы насчитывает лишь несколько десятилетий, вовсе не говорит о простоте и легкости этого пути.

Якутская литература никогда не отгораживалась непреодолимой стеной от художественных достижений других народов. Наоборот, постижение истории, окружающей среды и всего мира происходило за счет постоянного расширения своего культурного пространства. Якутская литература с самого рождения испытала благотворное влияние русской литературы. Первые национальные писатели якутского народа учились на лучших образцах русской классики. Эта учеба служила им стимулом собственного творческого роста.

Русско-якутский перевод имеет давнюю историю и огромное культурное значение для якутского народа. В отписках и подоб-

ных им документах первых воевод Якутского острога упоминаются первые устные переводчики – толмачи. Слово «толмач» произошло от тюркского «тыл мач» – “tilmaci” – «переводчик» [1]. Это слово из русского языка перешло в немецкий язык, так в немецком появилось слово “Dolmetscher”, что означает “(устный) переводчик; толмач” [2]. Отношение к толмачам, к сожалению, было неоднозначным, еще Петр I в своем знаменитом указе называл их «толмачи и прочая обозная сволочь» [3]. Однако, как пишет Е.И. Убрятова, ни одна серьезная экспедиция не обходилась без толмача [4]. О том, какое значение имели люди, знающие национальные языки, можно судить хотя бы по таким фактам: в документах толмач упоминается обычно рядом с начальником отряда [5], за отказ выполнять обязанности толмача люди, знающие туземный язык, несли суровое наказание [6]. В тех случаях, когда казаки злоупотребляли своей властью и занимались по существу грабежом, большую ответственность нес тот, кто знал язык данной народности и, наоборот, незнание языка служило в данном случае, смягчающим вину обстоятельством. Также Е.И. Убрятова отмечает, что толмачи по большей части были людьми неграмотными, никаких языковедческих работ они не писали. Но благодаря им в документы той эпохи вошло большое количество исконно якутских родовых названий, собственных имен, названий местности, а также предметов домашнего обихода туземного населения [7].

Знание якутского языка было также необходимым для всякого рода правительственных чиновников не только в начале присоединения, но и во все последующее время. Е.И. Убрятова сообщает, что даже и в конце XIX века все важнейшие официальные распоряжения и циркуляры (инструкции) обычно переводились на якутский язык, так как «язык якутов, по меткому выражению Островских, был международным на всем северо-востоке Сибири» [8]. Известный якутский тюрколог П.А. Слепцов в своей работе «Нууччалыы-сахалыы тылбаас историятыттан» («Из истории русско-якутского перевода») пишет, что история русско-якутского перевода началась в 1705 году, когда был опубликован анонимный перевод молитвы «Отче наш» на якутский язык [9]. Письменный перевод на якутский язык был необходим в первую очередь миссионерам. В отличие от толмачей они были людьми грамотными и оставили после себя большое количество всякого рода книг и

рукописей на якутском языке и о нем. Первая книга православного христианства на якутском языке «Зачатки вероучения» была издана в 1812 году в Иркутске. [10]

Э.К. Пекарский и Н.П. Попов в своей статье «Работы политических ссыльных по изучению якутской языка во второй половине XIX века» сообщают о том, что автором «Верхоянских сборников» Худяковым был составлен «Словарь якутского языка». Небольшой словарь якутских слов был у священника Орлова. Другим священнослужителем Ионовым был написан «Учебник якутского языка» для якутов, названный им «Олендорфия» (Олендорф – издатель немецких и французских учебников для русских). Из этой же статьи известно, что Пекарский составил небольшой словарь якутских слов. Он пользовался в работе словарем Натансона, Альбова и Орлова. Из центрального отдела Русского географического общества Пекарский смог получить «Якутско-русский словарь» П.Ф. Порядина и якутский текст «Верхоянского сборника» Худякова. [11]

Говоря о первых печатных опытах 1812–1821 гг., В.Н. Волкова в статье «Книга на языках коренных народов Сибири и Дальнего Востока в XIX–XX вв.» [12] отмечает интенсивное развитие переводческой деятельности православной миссии во второй половине XIX века. Автор приводит названия семи книг на якутском языке, отпечатанных московской Синодальной типографией в 1858 г. После создания в 1861 г. Якутской областной типографии издается еще 12 богослужебных, учебных книг на якутском языке, издания на эвенкийском, тунгусском языках.

Нужно отметить, что миссионерская письменность и перевод представляют собой обширный и очень интересный материал для исследования. Разница в качестве переводов, относящихся к разным периодам, отражает то, как изменялись со временем понимание адекватности и требования к качеству перевода религиозных текстов. Так, современный исследователь А.А. Васильева отмечает, что в начале своей деятельности члены Переводческой комиссии при составлении якутских текстов вынуждены были ориентироваться на разговорный язык, поскольку первые книги, относящиеся к началу XIX века, были написаны сложным, малопонятным для рядового якута языком, наблюдалась синтаксическая калька. Со временем качество якутских текстов улучшилось, и с середины XIX века переводимые тексты псалтырей становятся

образнее и доходчивее [13]. Появление миссионерского перевода имеет большое значение в создании грамматики якутского языка и якутской письменности. Именно миссионеры занялись составлением первых якутско-русских букварей, учебников русского языка и выработали новую транскрипцию якутского языка. По нашим сведениям в этот период было переведено около 60 книг.

Следующий период русско-якутского перевода охватывает конец XIX и середину XX вв. и связан с революционно-демократической деятельностью в Якутии и заставляет нас говорить о важной роли политических ссыльных в истории перевода Якутии. Огромное значение в росте классового самосознания якутского народа имела политическая ссылка. В Якутии отбывали ссылку представители трех поколений русских революционеров: декабристы, революционеры-разночинцы, в т.ч. вождь революционного движения 60–70-х гг. Н.Г. Чернышевский, затем народники и социал-демократы. [14] В начале XX в. в Якутске стали выходить общественно-политические газеты «Якутский край» (1907–08 гг.), «Якутская жизнь» (1908 г.), «Якутская мысль» (1909 г.), имевшие разделы на якутском языке, первый общественно-политический и литературно-художественный журнал на якутском языке «Саха саната» («Голос якута») (1912 г.) [15].

Как отмечено в статье Г.М. Васильева «Развитие местной печати», одним из существенных достижений газеты «Якутский край» является созданная при ней особая литературная страница на якутском языке под заголовком «Саха дойдута» («Якутский край»). В этом якутском отделе печатались главным образом статьи оппозиционного характера. Пользуясь тем, что представители местных властей не умели читать по-якутски, якутский отдел иногда печатал статьи, бичующие самодержавный строй. При газете «Якутская жизнь» существовал якутский отдел, носивший название «Саха оло5о» («Якутская жизнь»). [16]

После Октябрьской революции интерес к русско-якутскому переводу как к средству культурной революции возрос. Для ликвидации безграмотности и культурного просвещения стало выпускаться огромное количество переводной литературы. На страницах печатных изданий «Кыым» («Искра») и «Кы5ыл ыллык» («Красный путь»), как в то время назывался журнал «Хотугу сулус» («Полярная звезда», переименованный позже в «Чолбон» – «Венера»),

велась ожесточенная полемика по вопросам языка и литературы. В ней активное участие принимали известные якутские писатели А.Е. Кулаковский, А.И. Софронов, П.А. Ойунский, А.И. Иванов-Кюндэ, С.Р. Кулачиков – Элляй и многие другие. Так, А.И. Кюндэ опубликовал небольшую статью «Как нужно переводить», где тщательно проанализировал всего лишь два предложения из перевода одной работы В.И. Ленина, выполненного П.А. Ойунским. На основе этого анализа Кюндэ убедительно доказывает 3 вещи:

1. перевод должен основываться на достижениях сопоставительного изучения контактирующих языков;

2. перевод должен соответствовать нормам переводящего языка;

3. перевод должен передавать не только денотативное значение единицы, но и ее коннотативное значение.

Эти основополагающие принципы перевода были выведены Кюндэ задолго до становления переводоведения как отдельного раздела науки о языке. [17]

Таким образом, в этот рассмотренный нами период интерес к переводу возрастает. Впервые якутские писатели переводят произведения разного жанра: очерки, статьи, художественную литературу. В этот период, по нашим сведениям, вышло в свет более 80-ти переведенных произведений. Несмотря на большой размах сближения литератур, в указанный период традиция русско-якутского перевода, как явление или жанр, самостоятельно еще не оформилась.

В советские годы становится ясно, что переводчики начали интересоваться переводом не только как творчеством. Они пытались объяснить его успехи и неудачи не только уровнем своего таланта или вдохновением, но и с помощью научных методов литературоведения и переводоведения. Таковыми являются аналитические статьи С.Т. Руфова о переводах русской поэзии, С. Тумата. В этот период выходят переводы общественно-политических текстов, которые способствовали формированию официально-деловой разновидности публицистического и делового стиля якутского языка. Перевод с русского языка учебной и научно-популярной литературы, способствующий формированию учебно-педагогической, а также научно-популярного стиля якутского языка, представляет собой одну из переводческих проблем, связанную с терминологи-

ей. А также в этот период было положено начало развитию перевода с якутского языка на русский. Стали известны такие переводчики, как А. Ольхон, А. Гурулев, В. Державин.

Новый период русско-якутского перевода (1990-е гг.) связан с приданием якутскому языку статуса государственного. Добавим, и с объявлением английского языка – рабочим. В этот период резко возрастает количество переводимой литературы самых различных жанров и стилей. С 1980-х гг. начинается развитие перевода с якутского языка на иностранные и наоборот. В 1994 г. в Париже выходит перевод первой песни якутского эпоса П.А. Ойунского на французском языке под названием «Небесные воины страны якутов-саха». В 1993 г. на факультете иностранных языков Якутского государственного университета открыта кафедра перевода, а в 1999 г. – кафедра русско-якутского перевода при факультете якутской филологии и культуры. Перевод поэмы А.Е. Кулаковского «Сновидение шамана» (1999 г.) является одним из первых опытов перевода памятников якутской художественной литературы на английский язык. В 2002 г. вышло в свет письмо «Якутской интеллигенции» А.Е. Кулаковского в английском переводе. С 2005–2015 г. выходят в свет мультилингвальные издания олонхо П.Оготоева «Элэс Боотур» на русском, английском, корейском и французском языках. Переводятся на английский и французский языки киносценарии фильмов «Чингисхан» и «Глухой Вилюй», документальные фильмы по Якутии. В этот же период появляются и лексикографические работы: «Якутско-французский словарь», «Якутско-английский электронный словник», тематические глоссарии для разных отраслей профессиональной деятельности и др. В области науки, экономики, культуры начинает активно развиваться русско-английский перевод. В 2002 г. большой общественный и политический резонанс вызвал в республике авторизованный перевод монографии американского экономиста Джона Тихотского «Республика Саха (Якутия). Алмазная колония России». На английском языке выходят различные буклеты, справочники, фотоальбомы, CD-диски, веб-сайты. В октябре 2002 г. открыто Якутское региональное отделение СПР, объединившее переводчиков разных языков всей республики. Задачей данного творческого объединения является возрождение и развитие перевода с языков народов РФ на русский язык, а также перевод основных достиже-

ний науки, литературы, культуры народов Якутии на русский и иностранные языки. В 2013 г. происходит самое значимое событие в культурной жизни народа саха – издание английского перевода якутского героического эпоса олонхо П.А. Ойунского «Дьулуруйар Ньургун Боотур» под руководством А.А. Находкиной (Platon Oyunsky. (2014) Nurgun Botur the Swift. London).

Развивается устный перевод как отдельный вид переводческой деятельности. Развитие устного перевода как самостоятельной дисциплины диктует необходимость подготовки устных переводчиков со знанием китайского и якутского языка, особенно в сфере судопроизводства и государственных отношений. Этот период характеризуется резкой сменой вектора в сторону переводов на иностранные языки, что ведет к разным последствиям и выявляет проблемы развития перевода и связанных с этим проблем развития языков. Как отмечено выше, развивается лексикография, появляются различные глоссарии на иностранных языках, но 1) одновременно прекращается или сокращается издание переводов художественных произведений с якутского языка и других национальных языков РС(Я) на русский.

2) Российские издательства начинают выпускать только коммерческие переводы с иностранных языков, игнорируя сложившуюся в советское время традицию переводов с языков народов РФ.

3) Еще одна проблема – это низкая тарификация переводов с языков коренных малочисленных народов Севера и якутского языка по сравнению с переводами на иностранные языки. Огромная разница в оплате переводчиков с якутского языка на иностранные и на русский бросается в глаза. Для сравнения: перевод 1 страницы формата А-4 на иностранные языки стоит свыше 800 рублей за страницу плюс редактирование, которое оценивается отдельно, такой же перевод с якутского или языков коренных народов Севера стоит 5 руб. 95 копеек. Принятые в Якутском региональном отделении СПР «Единые переводческие тарифы» на деле применяются лишь к переводчикам на иностранные языки, в то время как переводчики с языков коренных малочисленных народов Севера и якутского языка получают крохи. Очевидно, что эти языки нуждаются в правовой и финансовой поддержке государства. Необходимо проведение единой государственной политики в сфере переводческих тарифов с языков коренных малочисленных наро-

дов Севера и якутского языка. Языки коренных малочисленных народов и якутский язык по статусу должны быть приравнены к иностранным языкам, а переводчики должны получать гонорары за свой труд не ниже, чем переводчики с иностранных языков.

4) Не обеспечивается право граждан, предусмотренное законом «О языках в РС(Я)» на письменный, устный последовательный и синхронный перевод с языков коренных малочисленных народов. Официальные мероприятия не сопровождаются последовательным или синхронным переводом на языки коренных малочисленных народов Севера и якутский язык и обратно, а телевизионные передачи, фильмы не обеспечены переводом с помощью субтитров.

Примечательно, что подобные проблемы – еще один яркий показатель развития переводческой деятельности в Якутии. Ведь, на этом этапе появляются переводы с якутского на иностранные языки текстов разнообразного содержания, что благотворно влияет на развитие функциональных стилей якутского языка – это и научно-публицистические, общественно-политические, информативные и художественные тексты. В этот период к переводу начинают относиться не только как к средству толкования информации, а как к большой и важной научной проблеме, от решения которой зависит судьба родного языка.

ЛИТЕРАТУРА

1. Древнетюркский словарь. / Под ред. Наделяева В.М., Насилова Д.М. – Л., 1969.

2. Большой немецко-русский словарь. / Под ред. Лепинга Е.И., Страхова Н.П. – М.: Русский язык, 1997, Т. I.

3. Виссон Л. Синхронный перевод с русского на английский. – М.:Р. Валент, 2000 – С. 19.

4. Убрятова Е.И. Очерк истории якутского языка. – Якутск: Як. кн. изд-во, 1945. – С.7.

5. Колониальная политика Московского государства XVII века. Л., 1936. с. 96, док. 44.

6. там же, с. 2, док. 3.

7. Убрятова Е.И. Очерк истории якутского языка. – Якутск: Як. кн. изд-во, 1945. – С. 9.

8. Убрятова Е.И. Очерк истории якутского языка. – Якутск: Як. кн. изд-во, 1945. – С. 5–6.

9. Слепцов П.А. Становление общенациональных форм. Новосибирск: Наука, 1989. – С. 45.

10. Отчет Якутской Переводческой комиссии за 1914 г. – С. 3–4.

11. 100 лет Якутской ссылки. Сб. якутского землячества. // Под ред. Брагинского М.А. – М.: Наука, 1934. – С. 344–352.

12. Книга в автономных республиках, областях и округах Сибири и Дальнего Востока. Новосибирск: Наука, 1990. – С. 9–14.

13. Васильева А.А. Синтетические трансформации при переводе с русского языка на якутский язык (на примере атрибутивных конструкций). // Дисс. к.ф.н. – Якутск, 2002. – С.5.

14. История Якутии. /Под ред. Башариной З. К. – Якутск: Як. кн. изд-во, 1988. – С. 78.

15. там же, С. 98.

16. Васильев Г.М. Развитие местной печати в связи с революцией 1905–1907 гг. // Революционные события 1905–1907 гг. в Якутии. – Якутск: Як. кн. изд-во, 1965. – С. 95.

17. Иванов А.И. – Кюндэ. Кысыл ыллык (Красный путь), № 3–4. – Якутск, 1931. – С. 4.

RESEARCH OF PROBLEM OF THE SEMANTIC ANALYSIS AND SYNTHESIS OF PREPOSITIONS (POSTPOSITIONS) IN THE RUSSIAN-KAZAKH MACHINE TRANSLATION

Diana Rakhimova

Al-Farabi Kazakh National University
Almaty, Kazakhstan

Kazakh and Russian are very complicated languages that have many differences in lexical and syntactic structure. Development of the Russian-Kazakh machine translation system enables us to identify some difficulties. This article considers the problem of the semantic analysis of prepositions of the Russian language and synthesis them in Kazakh and vice versa. Comparing prepositions (postpositions) of two languages we defined the characteristic properties of them and described the rules of transformation based on lexical, syntactic and semantic analysis of sentences.

Введение

Предлог – это служебная часть речи, которые необходимы для связи слов в словосочетании. Предлоги выражают зависимость одних слов от других, и могут быть при существительных, местоимениях и числительных. Они имеют различные виды образования, структуры и смысловые значения (Рубашкин В.Ш).

При переводе роль предлогов русского языка в казахском языке выполняют аффиксы и вспомогательные слова. И если предлоги русского языка стоять перед определяемым словом, то в казахском языке после: *Уехал на два месяца – екі айға кетіп қалды. Работать до вечера- кешке дейін жұмыс істеу;* Вроде бы никаких проблем с переводом предложных связей, т.е. определяется предлог в тексте из словаря и при переводе должен будет расположен после определяемого слова. Но при разработке машинного перевода (МП) с русского на казахский язык столкнулись с многозначностью предлогов в контексте предложения. В разных лексических конструкциях и с разными падежами предлоги могут иметь разные значения. Например: *забыл на столе* (пространственное значение), *отлучился на минуту* (временное значение), *верить на слово* (значение образа действия). Выше предложный метод перевода может быть не достаточен для полноты текста, т.к. имеет только стуктурную характерность в синтаксическом анализе

и генерации предложении в машинном переводе. В данной работе будет рассмотрены предложные связи, которые имеют семантические свойства.

При разработке русско-казахского словаря предлоги были разделены на две группы:

1) однозначные – предлоги и словообразования у которых есть определенный точный перевод: *до завтра -ертеңге дейін, в течение года – жыл бойы, через одинь час – бір сағаттан кейін, накануне наурыза- наурыз қарсаңында*;

2) многозначные – предлоги и словообразования которые имеют несколько значении перевода: *в декабре – желтоқсанда, в домеүйдің ішінде. под вечер – кешке қарай, под стол – үстел астында.*

Синтаксическое и семантическое описание предлогов практически значимо как раздел машинного словаря. С этой точки зрения состав словарного описания предлогов должен определяться исключительно востребованной алгоритмами анализа функциональностью словаря.

1. Модель структуры предложений русского и казахского языков

Формальные модели синтаксиса простых предложений казахского языка ориентированы на выделение трех типов фразовых структур: субъектной фразовой структуры(SP), глагольной фразовой структуры(VP) и объектной фразовой структуры(OP). Основой субъектной фразовой структуры является подлежащее, основой глагольной фразовой структуры является глагол и основой объектной фразовой структуры является объект действия. ниже представлена формальные модели синтаксиса простых предложений казахского языка с использованием аппарата формальных грамматик ().

С использованием нотации Бэкуса формальная модель структуры синтаксиса предложений казахского и русского языков будет иметь следующий вид.

S::= <SP><OP><VP>|<SP><VP><OP>|<OP><SP><VP>|<OP><SP><VP>|

<VP><SP><OP>|<VP><OP><SP>

(Данное правило представляет всевозможные варианты структур на уровне введенных фразовых структур)

<SP>::=<N>|<Adj><SP>|<Num><SP>|<N><SP>
<SP>::=<SP><Conn><SP>
<VP>::=<V>|<Aux><VP>|<Adv><VP>
<OP>::=<N>|<Adj><OP>|<OP><Conn><OP>

Здесь Adj –прилагательное, Num – числительное, Conn – союзы, Aux – вспомогательные глаголы, Adv – наречие.

Для русского языка добавляется правило с учетом предлогов:
<OP>::= <Prep><OP>|

Вышеуказанная модель преобразований структур предложений русского языка в структуры предложений казахского языка и наоборот используются при создании системы МП. По этим моделям разработаны алгоритмы, и разработана программа генератора

2. Семантический анализ и синтез предлогов в системе машинного перевода

Семантическая интерпретация предлога должна рассматриваться как частный случай более общей задачи – задачи семантической интерпретации синтаксических связей. Если между словами (в общем случае – текстовыми элементами) *W1* и *W2* парсером обнаружена синтаксическая связь, ставится вопрос о ее семантическом свойстве. В случае предложной связи вопрос может быть сформулирован так же, но в этом случае речь идет о конструкции вида *W1---> P---> W2*, где *W1* – (возможный) синтаксический хозяин, *W2* – синтаксический слуга, а предлог *P* маркирует связь между *W1* и *W2*, которая и является объектом семантической связи.

Задача машинного понимания текста сводится к переводу с естественного языка на язык представления знаний (ЯПЗ), в котором точно описаны правила построения и правила вывода (Рубашкин).

Синтезом предлогов (если он однозначный) является его перевод и структурное преобразование на целевой язык. Отношение предлогов рассматривается некой смысловой связью между объектами, которую при синтаксическом и семантическом анализе преобразуется в ЯПЗ. При генерации на целевой язык информация будет считываться с промежуточного ЯПЗ. Для предложных связей аксиомами могут быть некий набор правил, по которым может быть разрешена многозначность предлогов в МП. К отдельным

экземплярам относятся все предложные связи и словосочетания с предлогами, которые не поддаются аксиомам и правилам отношении. Это может быть художественные выражения, литературное высказывание или фразеологизмы. И такие отдельные экземпляры имеет свой полный смысловой перевод на выходной язык.

Так как предлог является корневым элементом своей предложной группы, были выделены его прямые аргументы из соответствующего поддерева семантического дерева предложения. Для обобщения переводной формулы предлога его аргументы были заменены семантическими классами атрибутов и грамматическими признаками.

Надо учитывать что организация связей и словоформ казахского языка отлична от русского языка. При переводе на казахский язык в различных случаях можно использовать определенный вид синтеза. Предлоги и предложные связи в казахском языке преобразуется с помощью присоединении аффиксов к основе слова (вариант 1) и/или вспомогательного слова стоящие после определяемого (вариант 2).

$$<pw_i>::= <w_j>|<w_j\ w^*_{j+1}>$$

Где p – предлог, w_i – определяемое слово входного языка, w_j – генерируемое слово выходного языка, w^*_{j+1} – вспомогательное слово. Разработан полный анализ предлогов русского языка и их преобразование на казахский язык. В таблице 1 проиллюстрированы примеры преобразования предлогов на казахский язык.

Таблица 1

Структурное соответствие предлогов в русском и казахском языке

Представление предлогов в русском языке	Преобразование предлогов на казахский язык. (1 вариант)	Преобразование предлогов на казахский язык. (2 вариант)	Пример
На сущ(ед\м.р) + е	зат + жат.с.(да.де..)	Зат + тәуел үстінде	На стол + е

На сущ(ед\ж.р) + е	зат+жат.с.(да.де..)	Зат+тәуел үстінде	На книг+е
На сущ(ед\с.р) + е	зат+жат.с. (да.де..)	Зат+тәуел үстінде	На окн+е
На сущ(ед\ж.р) + у	зат+бар.с. (ға.ге..)	Зат+тәуел үстіне	На книг+у
На сущ(ед\м.р) + у	зат+жат(да.де..)		на берег+у
На сущ(ед\м.р)	зат+бар.с. (ға.ге..)	Зат+тәуел үстіне	на стол
На сущ(мн\ж.р) + ах	зат+көп+ жат.с. (да.де..)	Зат+ көп +тәуел үстінде	на книг+ах
На сущ(мн\м.р) + ах	зат+көп+ жат.с. (да.де..)	Зат+ көп +тәуел үстінде	На стол+ах
На сущ(мн\с.р) + ах(ях)	зат+көп+ жат.с. (да.де..)	Зат+ көп +тәуел үстінде	На окн+ах(ях)
У сущ(ед\с.р)+а(я)	Зат+жат.с. (та.те..)	–	У окн+ а(мор+я)
У сущ(ед\ж.р)+ы	Зат+жат.с. (та.те..)	–	У ват+ы
У сущ(мн\ж.р)	Зат+ көп+ жат.с. (та.те..)	–	У книг
У сущ(мн\м.р)+ов	Зат+ көп+ жат.с. (та.те..)	–	У стол+ов
У сущ(мн\с.р)	Зат+ көп+ жат.с. (та.те..)	–	У окон
У сущ(мн\с.р)+ей	Зат+ көп+ жат.с. (та.те..)	–	У мор+ей
...			

К примеру переведем следующие простое предложение: “*Книга лежит на столе*» данный пример с предлогом можно перевести

в двух вариантах «*кітап үстелде жатыр*» или «*кітап үстелдің үстінде жатыр*». Конечно, тот или иной вариант не ошибочен и может использован по усмотрению пользователя. Для многозначных предлогов метод трансформации предложных связей на казахский язык не всегда можно описать с помощью синтаксической маркировки. В данном случае значение надо выбирать по контексту предложения. Например «*я пришел под вечер*» предлог «*под*» обычно описывает место и переводится как «*астында*», но в данном смысловое значение указывает время выполнения действия и должно иметь следующий перевод – «*мен кешке* ***қарай*** *келдім*». Основываясь на исследованиях семантических отношениях и многозначных предлогов была создана сопоставительная семантико-синтаксическая структура предложных связей русского и казахского языка.

Рассмотрим примеры с многозначными предлогами и определим их семантические отношения:

в *декабр****е*** → *желтоқсан****да****;*

в *дом****е*** →*үй****дің ішінде****;*

под *вечер*→*кеш****ке қарай****;*

под *стол* → *үстел* ***астында****;*

В примерах с предлогом «*в*» мы видим, что по лексическим и морфологическим признакам отличить разницу трудно, т.к. определяемые слова являются одной частью речи (имя существительное) и имеют одинаковое окончание «*е*». Но по смысловому значению первый пример описывает время, а второй – место. В реализации морфологического анализа и синтеза машинного переводчика словари будут оснащены дополнительными грамматическими и семантическими свойствами. И будут маркированы в базе данных для удобства реализации.

На этапе семантического анализа текста (Тукеев У.А., Рахмиова Д.Р. 2012) будут определены семантические атрибуты определяемого слова с предлогом, с помощью которых будут применены семантические правила. Для слов «*декабрь*» и «*вечер*» будут определены семантический атрибут времени и для фраз с помощью семантических ролей (правил) правильно определено смысловое отношение и корректно сгенерировано на выходной язык. В таблице 2 проиллюстрированы некоторые варианты.

Таблица 2

Примеры многозначных предлогов русского языка и их преобразования на казахский язык

Структура предложенных связей в русском языке	Смысло-вое значение	Структуры преобразования предлогов на казахский язык
На <w_i>	место	w_j + жатыс септік жалғауы(да.де..)
		w_j + тәуелдік жалғау *үстінде* (көмектес сөз)
	время	w_j + барыс септік жалғауы (қа,ке...)
Под <w_i>	место	w_j + тәуелдік жалғау *астында* (көмектес сөз)
	время	w_j *қарай* (көмектес сөз)
За <w_i>	место	w_j +тәуелдік жалғау *артында* (көмектес сөз)
	время	w_j +жатыс септік жалғауы (та,те..)
	цель	w_j *үшін* (көмектес сөз)

Приведем пример практического применения:

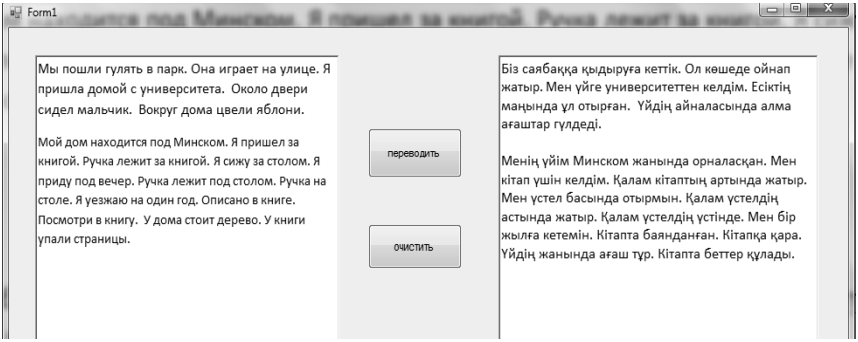


Рис. 1. Результат машинного перевода предложений с русского на казахский язык с многозначными предлогами

Как можно видеть на рисунке, было корректно определено смысловое значение предлогов и корректно подобрано соответствие на казахский язык.

Заключение

В данной работе была исследована проблема предлогов при машинном переводе. При анализе были учтены все возможные вариации слово изменений (часть речи, род, число, склонение, окончание и др.) в предложных связях, а так же были найдены и распознаны семантические свойства. Разработаны модель и алгоритмы анализа простых и многозначных предлогов русского языка с учетом характеристик синтеза на казахский язык. Реализована программа системы русско-казахского машинного перевода для простых предложений с предлогами.

ЛИТЕРАТУРА

Рубашкин В.Ш., *Семантической интерпретации предложных связей* [Электрон.ресурс].- url:http://www.dialog-21.ru/

Tukeyev U. Rakhimova D.R. (2012). *Augmented attribute grammar in meaning of natural languages sentences*. SCIS-ISIS 2012 The 6th International Conference on Soft Computing and Intelligent Systems. The 13th International Symposium on Advanced Intelligent Systems, (November 20–24), Kobe, Japan, 2012, – P. 1080–1084.

Рахимова Д.Р. (2014). *Построение семантических отношении в машинном переводе*. Вестник КазНУ, №1, 2014, – С. 90–101.

Тукеев У.А. *Разработка эффективных технологии компьютерного перевода казахского языка на английский и русский языки (и обратно) на основе методов формальных грамматик и статистических методов*: отчет о НИР (заключительный)/ ДГПНИИ ММ при КазНУ им аль-Фараби: рук. Тукеев У.А. Алматы, 2014. – 189 с. – № ГР 0112РК01467.

Jurafsky D., Martin J. *Speech and language processing: an introduction to nature language processing, computational linguistics, and speech recognition.* Pearson, Prentice hall. 2009, 988 p.

Кузнецов И.П., Сомин Н.В. (2010) *Особенности лексико-морфологического анализа при извлечении информационных объектов и связей из текстов естественного языка* [Электрон. ресурс]. – URL: http://www.dialog-21.ru/digests/dialog2010/materials/html/40.html.

EXPERIENCE OF CREATION OF TATAR-RUSSIAN STATISTICAL MACHINE TRANSLATION IN YANDEX

Andrey Sokolov[1], Andrey Egorov, Sergey Gubanov, Dmitriy Khristich, Mariya Schmatova, Irina Galinskaya, Alexey Baytin

Yandex, Moscow, Russia
[1]an-sok@yandex-team.ru

This paper describes an experiment on creation of a Tatar-Russian machine translation system based on Yandex statistical machine translation technology. The absence of large amounts of lexical data required for building translation models, as well as rich morphology of the Turkic languages, create serious problems for statistical machine translation. In particular, in the translation from Tatar to Russian many word forms remain untranslated. Therefore, in addition to statistical models, in this paper we used a morphological analyzer of the Tatar language, developed with the help from the Academy of Sciences of the Republic of Tatarstan. Experiments with morphological rules and disambiguation have helped to reduce the number of untranslated word forms and allowed achieving a reasonable translation quality on a rather small Tatar-Russian parallel corpus. As far as we know, it is the first attempt to build a statistical machine translation system for the Tatar language.

1. Введение

На татарском языке говорят более 5 миллионов человек. По количеству носителей языка татарский является вторым языком Российской Федерации после русского. В Республике Татарстан он является государственным, на нем ведется преподавание в школах, выпускаются телепрограммы, издаются газеты, публикуются сайты. Республика поддерживает международные связи со многими странами мира, молодежь в вузах активно изучает иностранные языки. Потребность в машинном переводе для татарского языка не вызывает сомнений, и по мере роста качества область применения перевода будет расти.

Инициатива проекта по созданию системы татарско-русского машинного перевода принадлежит сотрудникам НИИ «Прикладная семиотика» Академии наук Республики Татарстан, предложившим использовать для этого технологии статистического машинного перевода компании Яндекс. Необходимо отметить, что задача создания статистического машинного перевода для татар-

ского языка является сложной по многим причинам. Во-первых, количество электронных документов, необходимых для построения моделей перевода, на татарском языке пока еще очень мало. Во-вторых, татарский язык обладает богатой морфологией, и огромное разнообразие словоформ приводит к дефициту переводов для многих их них. В-третьих, и это является самой серьезной проблемой для машинного перевода, синтаксис татарского языка отличается от синтаксиса русского языка.

Создание корпусов текстов на татарском языке является важнейшей задачей проекта, однако для этого потребуется очень много времени, и ожидать появления больших параллельных корпусов в ближайшее время не приходится. Поэтому на первом этапе для снятия проблем лексической разреженности неоднозначности решено было заняться разработкой морфологического анализатора татарского языка. Большую помощь в этой работе нам оказали сотрудники Академии наук Республики Татарстан. Они предоставили данные и программы, позволившие существенно сократить время разработки.

В результате запуска морфологического анализатора удалось свести к минимуму количество непереведенных слов и, таким образом, существенно повысить общее качество татарско-русского перевода. Добавление возможности делать далекие перестановки также привело к заметному росту качества перевода.

2. Разработка экспериментальной версии татарско-русского статистического перевода

Разработка системы статистического машинного перевода для татарского языка включает в себя следующие компоненты:

- сбор параллельных текстов для базовой статистической модели;
- разработка морфологического анализатора;
- статистические перестановки.

2.1. Данные для базовой версии переводчика

Для статистического машинного перевода (Brown et al., 1990) необходим большой корпус параллельных текстов. Как и для всех остальных языков (для которых в Яндексе есть машинный пере-

вод), параллельные русско-татарские документы были собраны в интернете по технологии, похожей на метод, предложенный (Uszkoreit et al., 2010). Основным источником параллельных русско-татарских документов стали сайты государственных органов и СМИ Республики Татарстан, а также сайт Всемирного конгресса татар и несколько лингвистически-ориентированных ресурсов (например, информационный сайт о произношении слов на различных языках forvo.com). К сожалению, объемы татарских текстов в сети несопоставимы с объемами текстов для крупных европейских языков, поэтому начальное качество базового статистического перевода получилось низким.

2.2. Морфологический анализатор для тюркских языков

Типологически тюркские языки относятся к агглютинативным языкам, то есть аффиксы в этих языках «приклеиваются» к основе в определенном порядке не пересекаясь. Теоретически цепочка приклеенных аффиксов может быть бесконечной, а каждый последующий приклеенный аффикс каким-либо образом модифицирует смысл предшествующей ему цепочки морфем (на практике число приклеенных аффиксов обычно не превышает 2–3).

Такая особенность тюркских языков делает привлекательным описание их морфологии с помощью регулярных выражений. Регулярные выражения, в свою очередь, могут быть реализованы при помощи конечных автоматов.

У словоформы существует два представления – поверхностное и лексическое. Лексическим представлением аффикса называется идентификатор его грамматической роли, поверхностным – конкретная последовательность букв в конкретной словоформе.

Лексическое представление аффикса обычно кодируется некоторым мультисимволом, например, мультисимвол CaseGen служит для обозначения притяжательного падежа. Поверхностное представление аффикса может иметь несколько форм, выбор которых осуществляется по правилам сингармонизма в зависимости от соседних морфем. Например, для притяжательного падежа форма аффикса в турецком языке может быть: *nun, un, nün, ün, nin, in, nın* или *ın*. В татарском языке тот же аффикс представлен формами: *ның* или *нең*.

Таким образом, регулярными выражениями может быть описано как лексическое, так и поверхностное представление. В качестве символов регулярного выражения выступают аффиксы и основы слов. Допустимые комбинации основы и аффиксов задаются правилами морфотактики.

Регулярные выражения можно взаимно-однозначно описать конечными автоматами. Конечный автомат принимает последовательность входных символов (например, основа и лексическое представление аффиксов) и выдает последовательность выходных символов (основа плюс поверхностные представления аффиксов).

Если бы перед нами стояла только задача синтеза поверхностных форм из лексического представления, нам было бы достаточно программы, реализующей конечный автомат, и набора данных о морфотактике и сингармонизме конкретного тюркского языка.

Но перед нами стоит и обратная задача: по поверхностной форме получить лексическую. Для этого используется двунаправленный конечный автомат, называемый также трансдьюсером. Здесь нам не требуется писать правила обратного преобразования из поверхностной формы в лексическую. Для этого трансдьюсер использует правила прямого преобразования и инвертирует их, меняя направление работы.

Существуют различные формализмы представления данных для трансдьюсеров. Наиболее подходящим оказался формализм трансдьюсера *Xerox (xfst interface)* и его свободно распространяемый вариант *Foma*.

Для разработки татарского анализатора были использованы словарные униграммы, собранные по веб-документам, статьи с описаниями морфологий (Oflazer, 1994; Сулейманов и др., 1998, 2003) и трансдьюсер в формализме PC-KIMMO (Гильмуллин, 2009).

Отличие разрабатываемого трансдьюсера от классического состоит в том, что, во-первых, он разбирает словоформы, ошибочно написанные с нарушением сингармонизма, с неправильным использованием капитализации или диакритики, а во-вторых, умеет разбирать неизвестные слова. Также был увеличен объем используемого в трансдьюсере словаря.

2.3. Снятие морфологической неоднозначности

Задача снятия морфологической неоднозначности (*дизамбигуация*) состоит в выборе правильного морфологического разбора каждого слова в предложении с учётом его контекста. Разборы генерирует морфологический анализатор (см. раздел 2.2). В татарском языке примерно треть слов неоднозначны и имеют несколько разборов. Например, для слова *туры*, в зависимости от контекста, возможны следующие разборы:

- тур+Noun+3POSS_Sing (*место*, *время*, *момент <действия>*);
- туры+Adj (*прямой*, *прямо*).

Из приведённого примера видно, что *ы* может быть как частью леммы, так и аффиксом посессивности.

Морфологический разбор с дизамбигуацией – полезный шаг в анализе любого языка. В частности, для статистического машинного перевода дизамбигуация позволяет однозначно выбрать, какое из значений слова надо переводить (в примере выше – *место* или *прямой*).

Рассмотрим подробнее построение модели дизамбигуации. Пусть имеется обучающий корпус, в котором представлены эталонные морфологические разборы каждого слова в предложении. Для такого корпуса можно построить две независимые *N*-граммные языковые модели: по леммам l_i и по морфологическим признакам (*тегам*) t_i. Тогда вероятность разбора некоторого предложения по построенной модели будем оценивать по формуле (1).

$$P(l_1t_1, \ldots, l_kt_k) = P_l(l_1, \ldots, l_k) \cdot P_t(t_1, \ldots, t_k) \quad (1)$$

где $l_1 \ldots l_k$ – последовательность лемм в предложении;
$t_1 \ldots t_k$ – последовательность цепочек тегов в предложении.

В качестве P_l и P_t будем использовать *N*-граммные языковые модели со *Stupid Backoff* сглаживанием (Brants et al., 2007), в которых вероятность последовательности $(w_1 \ldots w_k)$ задается по формуле.

$$P(w_1 \ldots w_k) = \prod_{i=1}^{k} P(w_i | w_{i-N+1} \ldots w_{i-1}) = \begin{cases} \dfrac{c(w_{i-N+1} \ldots w_i)}{c(w_{i-N+1} \ldots w_{i-1})}, \text{если } c(w_{i-N+1} \ldots w_i) > 0 \\ 0.5 \cdot P(w_i | w_{i-N+2} \ldots w_{i-1}), \text{иначе} \end{cases} \quad (2)$$

где $c(w_1 \ldots w_n)$ – частота *n*-граммы (леммы или цепочки тегов соответственно) в обучающем корпусе.

Выбор наилучшего разбора предложения состоит из следующих этапов:

- получение всех возможных разборов для каждого слова в предложении;
- построение графа-решётки для полученных разборов отдельных слов;
- выбор наилучшего пути в графе-решётке.

Рассмотрим данный процесс на примере разбора следующего предложения на татарском языке:

Хәзер Мәскәүдә яхшы консультантлар.

(В Москве сейчас хорошие консультанты.)

Опустив процесс получения возможных разборов для каждого слова, приведём граф-решётку для предложения (см. рис. 1).

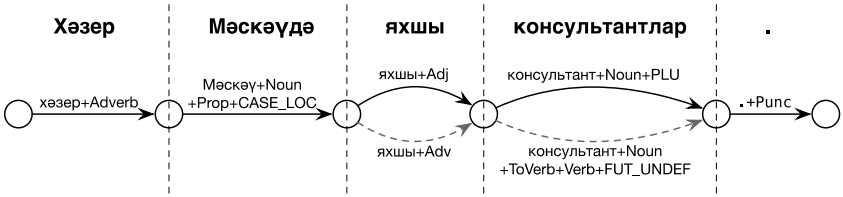


Рисунок 1. Граф-решётка для разборов исходного предложения; сплошной линией выделен наиболее вероятный разбор.

Представленный граф является взвешенным, при этом вес каждого ребра определяется по модели дизамбигуации (см. формулу (2)). Каждый путь в графе соответствует одному из разборов предложения. Поиск наиболее вероятного разбора осуществляется при помощи алгоритма Витерби (Viterbi, 1967).

Описанная в данном разделе модель дизамбигуации требует наличия размеченного обучающего корпуса, когда каждой словоформе в предложении соответствует единственный разбор. Для татарского языка такого корпуса не оказалось, из-за чего было предложено построить модель без учителя.

В случае, когда предложение имеет p возможных разборов, разумным является предположение о равной вероятности каждого из них. Для каждого из этих разборов мы порождаем n-граммы (по

леммам и тегам), умножая частоты на `1/p`. Получившиеся дробные частоты можно напрямую использовать в формуле (2).

2.4. Морфологический фильтр

По результатам работы трансдьюсера и снятия морфологической неоднозначности было получено представление слов вида «основа+морфологическая информация». Однако, чтобы морфология эффективно работала на этапе выравнивания, необходимо отрезать цепочки тэгов от основы. В данной работе были использованы «базовые» правила отрезания цепочек тэгов: слово разделяется на две части – основу и цепочку тэгов (без разделения цепочки тэгов на части). Таким образом, цепочки тэгов стали самостоятельными токенами на этапе пословного выравнивания. Тэги, приписанные знакам препинания, были удалены.

Так, предложение, представленное на рис. 1, по результатам работы фильтра принимает вид:

Хәзер$ $Adverb Мәскәү$ $Noun+Prop+CASE_LOC яхшы$ $Adj консультант$ $Noun+PLU.

2.5. Статистические перестановки

Поскольку порядок слов в исходном и переведенном предложениях зачастую сильно отличается (особенно для далеких языков), монотонного алгоритма перевода, рассматривающего слова по очереди слева направо, оказывается недостаточно. В то же время, переводить слова в произвольном порядке тоже нельзя, поскольку это, во-первых, искажает смысл и в целом негативно сказывается на качестве перевода, и во-вторых, сильно увеличивает вычислительную сложность задачи. В таких случаях приходится использовать т.н. модель перестановок, которая ограничивает множество допустимых перестановок слов и вводит штрафы за их использование.

Рассмотрим одну из наиболее простых и широко используемых моделей перестановок – линейную (Al-Onaizan et al. 2006). Штраф за перестановку слов в этой модели пропорционален величине т.н. *distortion*'а – количества исходных слов, пропущенных при переводе очередного фрагмента предложения. Общий штраф за перестановки для данного варианта перевода в такой модели равен:

$$D(e, f) = -\sum_i d_i, \quad (3)$$

где e – исходное предложение, f – его перевод, а d_i определяется как abs(индекс последнего слова (i-1)-го переведенного фрагмента + 1 – индекс первого слова i-го фрагмента).

Важным параметром моделей, использующих *distortion*, является т.н. *distortion limit*, то есть максимальное допустимое значение *distortion* для перевода. Низкие значения этого параметра ускоряют работу переводчика, но не допускают далеких перестановок, слишком высокие же – наоборот, сильно замедляют и могут приводить к ухудшению качества.

3. Эксперименты

3.1. Данные для экспериментов

Для проведения экспериментов был использован корпус параллельных веб-документов, содержащий несколько сотен тысяч параллельных предложений.

Для оценки качества перевода использовался тестовый набор, составленный из 400 предложений новостной тематики на русском языке, переведенных на татарский язык специалистами АН РТ.

3.2. Метрики

Для оценки качества машинного перевода традиционно используется метрика BLEU (Bilingual Evaluation Understudy), оценивающая близость предложения, переведенного системой машинного перевода к переводу эксперта-переводчика (Papineni et al., 2001). Метрика определяет процент N-грамм, совпавших в машинном и эталонном переводах предложения.

Помимо основной метрики BLEU при оценке качества перевода оценивался пословный OOV (Out Of Vocabulary) – процент не переведенных системой машинного перевода слов. В общем случае чем меньше OOV, тем выше качество машинного перевода.

3.3. Результаты экспериментов

Чтобы посмотреть, как влияют описанные выше методы на качество татарско-русского перевода, было проведено несколько экспериментов. Список экспериментов приведен ниже:

- BASE – базовая статистическая модель;
- MORPH – к базовой статистической модели добавлен морфологический фильтр, описанный в разделе 2.4;
- MORPH-D – к базовой статистической модели добавлен морфологический фильтр, толерантный к диакритике, описанный в разделе 2.4;
- REORD – к морфологической модели (MORPH-D) добавлены статистические перестановки, описанные в разделе 2.5. В эксперименте значение *distortion limit* было равно16.

Результаты экспериментов представлены в Таблице 1.

Таблица 1

Результаты экспериментов

	BASE	*MORPH*	*MORPH-D*	*REORD*
BLEU	9.35	8.5	8.86	10.41
OOV	15.3	7.7	7.7	6.5

Таблица 2

Примеры переводов разных версий системы

Source	*Reference*	*BASE*	*MORPH*	*MORPH-D*	*REORD*
Дәфтәр	Тетрадь	Тетрадь	Книга	Книга	Книга
Тал	Ива	Тал	Без	Без	Без
Сабын	Мыло	Мыльным	Сабын	Мыла	Мыла
Алмагач	Яблоня	Яблони	Яблони	Яблони	Яблони
Кабымлык	Закуска	Кабымлык	Хотя	Кабымлык	Кабымлык
Бүген музейга кил	Приходи сегодня в музей	Сегодня приходи	Сегодня в музей пришли	Сегодня в музей пришли	Сегодня пришел в музей
Балалар су коена	Дети купаются	Дети воды коена	Дети купаться можно	Детям можно купаться	Дети могут купаться

С добавлением морфологического фильтра значение BLEU упало на 0.5, однако в два раза уменьшился процент непереве-

денных слов (OOV), а значит перевод стал более осмысленным. Например, базовый статистический перевод (BASE) не справился с переводом предложения “Балалар су коена” (см. Таблицу 2), а при добавлении морфологического фильтра (MORPH), хотя перевод получается не очень гладким, мы уже можем догадаться о смысле исходного предложения. При использовании трансдьюсера, толерантного к диакритике (MORPH-D), слово сабын стало переводиться правильным значением (в отличие от эксперимента с обычным трансдьюсером (MORPH)).

Заметный прирост по BLEU мы получили, добавив модель перестановок (REORD). Перевод также стал более гладким и читаемым.

4. Заключение

В данной работе мы поделились опытом создания первой из известных нам систем статистического машинного перевода для татарского языка. Из-за крайне малого объема доступных для обучения моделей данных и, даже в большей степени, из-за лингвистической сложности татарского языка проект получился сложным и, по сравнению со многими другими языками, весьма трудоемким. Собранный по обычной схеме статистический татарско-русский перевод оказался ожидаемо непригодным для использования в силу большого количества непереводимых слов. Перевод, получившийся после подключения морфологического анализатора, остался по-прежнему очень далеким от желаемого, но уже читаемым, т.е. ранее непереводимые слова стали переводиться, что дало возможность понимать смысл предложений.

Работы по улучшению татарско-русского статистического машинного перевода планируется продолжить. Мы также надеемся на рост качества перевода за счет обратной связи, возникающей по мере увеличения количества пользователей татарского перевода.

Благодарности

Авторы выражают глубокую благодарность Д.Ш. Сулейманову, Р.А.Гильмуллину и другим сотрудникам НИИ «Прикладная семиотика» Академии наук Республики Татарстан за инициативу и техническую поддержку.

ЛИТЕРАТУРА

Oflazer Kemal (1994) Two-level Description of Turkish Morphology, Literary and Linguistic *Computing*, Vol: 9, No: 2.

Сулейманов Д.Ш., Гильмуллин А.А., Гильмуллин Р.А. (1998). Файл фонологических правил как основа двухуровневого морфологического анализатора татарского языка. *Телеконференция*.

Сулейманов Д. Ш., Гатиатуллин А.Р. (2003). Структурно-функциональная компьютерная модель татарских морфем. Казань: *Фэн.*

Гильмуллин Р. А. (2009) Математическое моделирование в многоязыковых системах обработки данных на основе автоматов конечных состояний. *Автореферат диссертации на соискание ученой степени кандидата физико-математических наук.* Казань.

Brants, T., Popat, A.C., Xu, P., Och, F.J., and Dean J. (2007). Large Language Models in Machine Translation. *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* Prague. pp. 858–867.

Viterbi, A.J. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, Vol. 13, No. 2, April pp. 260–269.

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*. pp. 311–318.

Brown P., John Cocke, S. Della Pietra, V. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, P. Roossin (1990). A statistical approach to machine translation. *Computational Linguistics (MIT Press)* 16 (2): 79–85.

Uszkoreit, J., Ponte, J. M., Popat, A. C., & Dubiner, M. (2010). Large scale parallel document mining for machine translation. *In Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics*. pp. 1101–1109.

Al-Onaizan, Y. and Papineni, K. (2006). Distortion models for statistical machine translation. *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics,* pp. 529–536, Sydney, Australia. Association for Computational Linguistics.

A FREE/OPEN-SOURCE MACHINE TRANSLATION SYSTEM FROM ENGLISH TO KAZAKH

Aida Sundetova[1], Mikel Forcada, Francis Tyers

Scientific Research Institute of Mathematics and Mechanics, Al-Farabi Kazakh National University, Al-Farabi av., 71, Almaty, 050040, Kazakhstan
Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Alacant, E-03071, Spain
Gielladiehtaga instituhtta, UiT Norgga árktalaš universitehta. Romsa, N-9037, Norway
[1]sun27aida@gmail.com

This paper presents the current state of development a shallow-transfer rule-based machine translation (MT) system from English to Kazakh. The main syntactic and morphological differences between the two languages are presented: Kazakh language shows clear ordering of morphemes and they have complex phonological changes, which depend on neighboring morphemes and such interactions are called sonorization, vowel harmony, etc., whereas English is morphologically not too complex as Kazakh language; syntactically, between English and Kazakh, there are many differences, for instance, in order of members of sentence: subject–object–verb order (compare with subject–verb–object in English), using prepositions in English, whereas in Kazakh it is transformed into postpositions, lack of definite articles (extensively used in English).

In this paper is showed how the machine translation system was designed to tackle these challenges. Machine translation system is build on Apertium free/open-source machine translation platform and there is shown the structure of this system and how it works. For English-Kazakh language pair there were developed linguistic data such like monolingual (Kazakh, English), bilingual dictionaries (English-Kazakh), lexical-selection, constraint grammar and structural transfer rules. We described structures and building features of each dictionary and rule. For instance, to create Kazakh morphology (monolingual dictionary) is used the Helsinki Finite State Toolkit, which implements finite-state morphological transducer, which could perform agglutination of Turkic languages, and, for English, monolingual dictionary is built with paradigms and lemmas from each form of word. We show example of translations, an evaluation of system coverage and translation quality and outline and future work.

1. Introduction

English language is a West Germanic language and Kazakh belong to group of Turkic languages. Therefore, the Kazakh, as most Turkic

languages, has a very rich morphology and shows agglutination, whereas English has a very simple morphology.

Furthermore, there are more differences between the syntax of Turkic languages and English: head-final syntax with modifiers and specifiers always preceding the modified/specified (normally following in English), overt case marking allowing for a rather free ordering of arguments (versus a more fixed order in English), verbal-noun-centered structures where English uses modal verbs (must, have to, want to) or verbal-noun or verbal-adjective-centered constructions where English has subordinate clauses using finite verbs with relatives or subordinating conjunctions (the book which I read, the place where I saw him, before he came), lack of a parallel of the English verb have, as used for possession, etc. For an account (in Russian) of syntax differences between English and Kazakh (Pecherskikh and Amangeldina, 2012).

This paper describes work in progress of development of machine translation system for English-to-Kazakh, which developed by using Apertium free/open-source machine translation platform (Forcada et al., 2011, http://www.apertium.org).

2. The Apertium platform

Apertium is a free/open-source rule-based machine translation (MT) platform that was built in 2005 by the Universitat d'Alacant. At first, it was initially aimed to translate texts between closely related languages, then it was extended to deal with unrelated languages. This platform has next components: machine translation engine, developer's tools, and linguistic data for an increasing number of language pairs and they are licensed under the free/open-source GNU General Public License (GPL, versions 2 and 3).

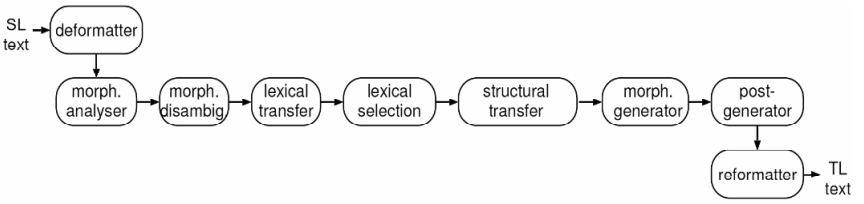


Fig. 1. The pipeline architecture of the Apertium system

- **De-formatter.** Separates the text to be translated from the formatting tags. Formatting tags are encapsulated in brackets so they are treated as "superblanks" that are placed between words in such a way that the remaining modules see them as regular blanks.

- **Morphological analyser.** For each surface form (SF) the morphological analyser generates one or more lexical forms (LF), which consist of: lemma (dictionary or citation form), lexical category (or part-of-speech), and inflection information.

- **Part-of-speech (POS) tagger.** This module chooses one of the LFs of an ambiguous SF.

- **Lexical transfer.** It reads each source-language LF and delivers the corresponding target-language LFs. This module uses a bilingual dictionary. Multiword lexical units can be translated as a single LF.

- **Lexical selection.** It uses rules to select one of the target-language LFs, as described in (Tyers et. al., 2012).

- **Structural transfer**. This module uses pattern matching to identify sequences of LFs (phrases or segments), which need syntactical processing. It uses files with rules, which specify syntactic transformations such as word reorderings, lexical changes such as changes in prepositions, and agreement between target language lexical forms.

- **Morphological generator.** It transforms the sequence of target-language LFs, produced by the structural transfer, to a corresponding sequence of target-language SFs.

- **Post-generator.** Performs some minor orthographical operations in the target language.

- **Reformatter.** It places format tags back into the text so that its format is preserved.

3. Linguistic data

Apertium uses text-based (mainly XML-based) formats for linguistic data, which include bilingual and monolingual dictionaries, structural transfer and lexical selection rules.

1.1. Dictionaries

Dictionaries are used in lexical processing: monolingual dictionaries for morphological analysis of English and generation of Kazakh and bilingual dictionaries for English–Kazakh lexical transfer.

Morphological dictionaries are used to define the correspondences between LFs and SFs; they contain a definition of the source-language alphabet and grammatical categories and attributes (such as noun, verb, plural, locative, etc.), a list of all lexical units, and inflection paradigms for all lexical categories; paradigms describe regularities in the correspondences between parts of SFs and LFs.

The English dictionary comes from existing data in Apertium, such as the English–Spanish language pair.

The Kazakh dictionary (Salimzyanov et al., 2013) is turned into a finite-state morphological transducer, using the Helsinki Finite State Toolkit (Linden et. al., 2011). The dictionary is based on two-level rules, and uses the lexc formalism for defining lexicons through word classes and subclasses, and the twol formalism for morphophonological rules such as vowel harmony, desonorization, nasalization, etc. An example of morphophonological rule: depending on the preceding vowels and consonants, the plural suffix -LAr[1] for noun could become -тар, -тер, -лар, -лер, -дар, -лер: кітап+тар, мектеп+тер, etc.

The Englis–Kazakh bilingual dictionary provides correspondences between English LFs and Kazakh LFs. Ambiguity is allowed: one of the LFs will be chosen by lexical-selection rules depending on context (see below). This dictionary currently contains 13,135 stems (Sundetova and Karibaeva, 2013).

1.2. Rules

1.1.1. Disambiguation rules

The problem of part-of-speech (PoS) ambiguity is solved using the Constraint Grammar (CG) (Karlsson, 1995) formalism: CG rules choose one of the LFs obtained for each SF.

---

[1] Uppercase Latin letters are used for archiphonemes (actually archigraphemes) that are realised as phonemes (actually graphemes) after morphophonological rules have been applied.

1.1.2. Lexical selection rules

Lexical-selection rules choose one of the alternative target-language LFs corresponding to one source-language LF, as described by (Tyers et. al., 2012). Alternative translations are defined in bilingual dictionary by multiple entries for each source-language LF. For example, the adjective beautiful could be translated as сұлу, if the following noun is a person: beautiful girl→сұлу қыз; or as әдемі or көркем, if the following noun means place: beautiful mountains→әдемі таулар. Table 1 shows examples of lexical-selection rules:

Table 1

Example lexical-selection rules

SL word w	TL word	Context	Example
residence	Мекен	default	(I live in) residence (Мен) мекенде өмір сүремін
	Резиденция	*w*_of_president	(I see the) residence of president (Мен) президенттің резиденциясын көремін
boot	Бәтеңке	default	I bought boots Мен бәтеңкелерді сатып алдым
	Жүксалғыш	*w*_of_car	(He opened a) boot of the car (Ол) машинаның жүксалғышын ашты
anything	Бір нәрсе	default	I see anything Мен бір нәрсе көремін
	Еш нәрсе	not_[verb]	(I do) not do anything (Мен) еш нәрсе істемеймін

1.1.3. Transfer rules

For English–Kazakh transfer is performed in three stages (Sundetova et. al., 2013):

- A first round of transformations (“chunker”) detects source language (SL) LF patterns and generates the corresponding sequences of

target language (TL) LFs grouped in chunks representing simple constituents such as noun phrases, prepositional phrases, etc.

- The second round ("interchunk") reads patterns of chunks and produces a new sequence of chunks. This is the module where one can attempt to perform some longer-range reordering operations, agreement between chunks, case selection, etc.

- The third round ("postchunk") transfers chunk-level tags to the lexical forms they contain and whose lexical-form-level tags are linked (through a referencing systems) to chunk-level tags (for instance, case determined for a noun phrase is transferred to the main noun), and removes all grouping information to generate the desired sequence of TL LFs.

The structural transfer module in Apertium processes the stream of source-language lexical form – target-language lexical form pairs (SL LF–TL LF pairs) and transforms it into a sequence of TL LFs after a series of structural transfer operations specified in a set of rules: reordering, elimination or insertion of TL LFs, agreement, etc.

This section describes the current structural transfer in apertium-eng-kaz, except work from (Sundetova et. al., 2013). English–Kazakh chunker rules, interchunk rules and an additional clean-up stage will be described in the following 3 sections.

1.1.3.1. Chunker

In the first round of structural transfer, rules segment sentences into chunks, such as short noun phrases, adjective phrases, verb phrases and adpositional phrases (that is, prepositional phrases in English and postpositional phrases in Kazakh). Chunking rules, of which there are currently 168, identify 8 kinds of chunks and translate them into equivalent Kazakh chunks, leaving some adaptations to be performed in later stages of structural transfer (for instance, the morphological case of noun phrases).

- Noun phrases (NP): general noun phrases consist of noun plus adjective, numeral or prepositions. Unusual types of noun phrases consist of gerunds (*-ing* ending): I like playing – *Мен ойнау+ды*(accusative case) *жақсы көремін*(*I playing*-ACC *like*). As can be seen from example, gerunds could get case as simple noun phrase, also its possessive could be determined in next stages.

• Prepositional phrases (PP): English prepositional phrases are translated into Kazakh as postpositional phrases, there are three possible outcomes with different cases: genitive -*NIң*[1], in which will case the phrase will be marked GenP; locative -*{D}{A};*[2] ablative -{*D}{A}н*, etc.; using postpositional constructs based on positional nouns such as *астында* (‘under’), *үстінде* (‘on’), etc.

• Verb phrases (VP): Translation of English verb phrases into Kazakh is not always straightforward. For instance, tenses expressing continued activity, such as the English present continuous or past continuous (*I am playing*, *I was playing*), have to be detected and mapped onto sets of two lexical units (*Мен ойнап жатырмын*, *Мен ойнап отырдым*). Special types of verb phrases like pseudo verbs: like, hate, enjoy*, etc.* are used to detect pseudo verb + gerund construction: *I enjoy dancing* – Мен билеуді **ұ**натамын, where pseudo verb get number and person, not the second verb as in present continuous sentences; auxiliary question verb: *do/did?, be/was/were?*, etc. are detected to generate in *interchunk* stage question words *ма/ме/ба/бе* and determine which tense it is(see Table 2):

Table 2

Examples of translating questions

English tense	Example	Chunker analyse of verb phrase
Present Simple	*Do you play?*	VP_q<VPQ><aorist>{ }
Perfect	*Have you been?*	VP_qhave<VPQ><past>

• Adjectival phrases (AdjP): In Kazakh noun phrases, adjectives come before nouns and do not show any agreement with nouns. Adjectives can also appear in separate adjective phrases. Two kinds of adjective phrases are distinguished: AdjP (for isolated adjectives and comparative adjectives) and SupP (superlative adjectives).

---

[1] In the genitive ending -*{N}{I}ң*, the archiphoneme {N} may be realised as т, д, or н and the archiphoneme {I} may be realised as і or ы depending on the previous phonological context.

[2] *{D}* can be *{ə}* or *{m}*, and *{A}* can be *{e}* or *{a}*, depending on the phonological context.

1.1.3.2. Interchunk processing

The second round of structural transfer is currently performed by a proof-of-concept set of 140 rules, representative of following operations:

- Inter-chunk agreement between subject noun phrase and verb phrase in their person and number.
- Assigning case to noun phrases (which are generated without case by the chunker): for instance, accusative case for objects (*I bought the table → Мен үстелді сатып алдым*), genitive case for obligatory constructs (*I must see → менің көруім керек*), dative case for the verb to need (*I need a doctor →Маған дәрігер керек*), locative case for possession (*I have a book → Менде кітап бар*), etc.
- Reordering: placing of object before verb (I[1] bought[2] the table[3] → *Мен*[1] *үстелді*[3] *сатып* алдым[2]), placing of prepositional pharses before the verb (They[1] played[2] on top of the tree [3] → *Олар*[1] *ағаштың үстінде*[3] *ойнады* [2]), etc.
- Adding question particle *ма*/*ме*/*ба*/*бе* at the end of a question, if the question mark chunk "?" is detected (Did<VP_ques> you<NP> watch<VP> last film<NP> ?<Q_m>→ *Сіз*<NP> *соңғы фильмді*<NP> *көрдіңіз*<VP> *бе*<ques> ?<Q_m>).
- For sentences like "I am a student" and "He is from Kazakhstan" replacing auxiliary verbs (*am*/*was*/*were*/*is*) at the end and assigning number and person to the complement NP (*student – student<p1><sg>*) and complement PP (Kazakhstan<*p3*><*sg*>).
- Place possessive for long noun + noun + noun structures (The university of city of Kazakhstan – *Қазақстан қаласының университеті*).
- Changing verb tense to conditional (<cond>), if it comes after "if": *If I come<aor>→<cond>, I will go – Егер мен кел**се**м, мен барамын.*

The set of rules has to be extended, as many combinations of the above phenomena are still not covered.

1.1.3.3. Postchunk

English–Kazakh postchunk rules straightforwardly transfer chunk-level tags to the head word (for instance, if the noun phrase is in locative, the locative case is transferred to the head noun).

1.1.3.4. Postchunk and cleanup

An additional phase was added to English–Kazakh transfer to be able to remove or give default values to tags which cannot be determined during the chunker and interchunk steps. For instance, if a noun phrase was not determined to be acting as an object and therefore the head noun did not get accusative case, such a noun would be received with a <CD> ("case to be determined") tag, which after the cleanup would be set to the default value (nominative). Other operations carried out at this stage ensure the agreement between noun or adjective and the copula verb "е" (Мен дәрігер+е<p1 person singular> → Мен дәрігермін), or deciding the actual form of the question particle according to the preceding word (Сіз келдіңіз + ме? → Сіз келдіңіз + бе?).

4. Evaluation and results

The current system can translate simple sentences and questions. The English-Kazakh bilingual dictionary contains 13,135 entries. There are 168 "chunker" rules and 140 interchunk rules. Evaluation was performed on revision 60571 in the Apertium Subversion repository. Lexical coverage is calculated over EuroParl[1], SETimes[2], NewsCommentary[3], Wikipedia[4].

Table 3 presents the size of each corpus and the vocabulary coverage of the system for that corpus.

Table 3

Vocabulary coverage of the English--Kazakh system over four available corpora

Corpus	Tokens	Coverage (%)
SETimes	5.1 mil.	97.90%
NewsCommentary	6.5 mil.	96.27%
EuroParl	54.5 mil.	97.95%
Wikipedia	1.8 bil.	84.67%

[1] *http://www.statmt.org/europarl/v7/es-en.tgz*

[2] *http://nlp.ffzg.hr/data/corpora/setimes/setimes.en-tr.txt.tgz*

[3] *http://www.statmt.org/wmt13/training-parallel-nc-v8.tgz*

[4] *http://dumps.wikimedia.org/enwiki/20140402/enwiki-20140402-pages-articles.xml.bz2*

As can be seen from Table 3, the coverage of dictionaries is good, with an average score of 94%. Our system outperforms the other available RBMT system, but falls short of the state-of-the-art performance represented by Google Translate. However, unlike the state of the art, which uses corpora which are unavailable to the general public, our system and the linguistic resources used in it are free and open/source and are open to improvement.

The output of the system was evaluated with BLEU (Papineni et. al., 2002) and the word error rate metric (Levenshtein, 1966). We chose a short text in English, and by postediting output of English-Kazakh system, a parallel text was built. The output of each of the machine translation systems was postedited independently to avoid biasing in favour of one particular system (see Table 4).

Table 4

Metric results for the three systems compared

System	BLEU	WER
apertium-eng-kaz	44.23%	42.88%
Google	62%	25.92%
Sanasoft[1]	21%	74.52%

MT systems have some mistakes in translations, which have been shown with * (see Table 5). The Google MT system gets a higher BLEU score, but in translation of selected phrases, it makes common mistakes such as not assigning the right possessive and case. The Sanasoft system has more errors as regards the translation of words, and many out-of-vocabulary words.

Table 5

Qualitative evaluation

Structure	English	System	Translation
Noun phrases	My difficult exercises	Apertium	Менің қиын жаттығуларым
		Sanasoft	Менің қиын жаттықтырып *жатыр
		Google	Менің қиын *жаттығулар

[1] *http://www.sanasoft.kz/c/ru/node/47 (in Russian) http://www.sanasoft.kz/c/kk/node/53 (in Kazakh).*

Structure	English	System	Translation
	Conan Doyle	Apertium	*Дойлдың Конан
		Sanasoft	*Conan *Doyle
		Google	Конан Дойл
Prepositional phrases	Under three big trees	Apertium	үш үлкен ағаштың астында
		Sanasoft	*Three үлкен *ағаштар астында
		Google	үш үлкен *ағаштарды астында
Adjective phrases	The most beautiful	Apertium	ең әдемі
		Sanasoft	*Көпшілік әдемі
		Google	ең әдемі
Modal verbs	I must pay	Apertium	Менің төлеуім керек
		Sanasoft	Мен *must *pay
		Google	*Мен *төлеуі тиіс
	It must be James	Apertium	Бұл Джеймс *болады
		Sanasoft	*Оған Джеймс *must *болып *жатыр
		Google	Джеймс болуы тиіс
Question	Is it right?	Apertium	*Екен *дұрыстың ол?
		Sanasoft	Бұл *түзуі?
		Google	Бұл дұрыс па?

5. Conclusions

We have presented the design of a free/open-source rule-based MT system from English to Kazakh. The current English–Kazakh machine translation system already successfully translates noun-phrases, verb-phrases, prepositional-phrases, and adjectival-phrases, and contains a good vocabulary for testing purposes.

We plan to continue developing the English–Kazakh pair, aiming at extending the coverage to 98% of the reference corpora. Additionally, we will improve the quality of translation by adding more rules, such like constraint grammar rules, structural transfer and lexical-selection

rules. The future plan is to use the created data with other free/open-source MT systems involving Turkic languages or in systems having Kazakh as target language to make transfer systems between the Turkic or other language pairs. Related work is currently ongoing with Russian–Kazakh and Kazakh–English (Sundetova et. al., 2014).

Our system is available as free/open-source software and the whole system may be downloaded from SourceForge.[1]

Acknowledgements

MLF thanks the Kazakh state program for the attraction of foreign scholars and Prof. Ualsher Tukeyev for supporting his visit to the Kazakh National University, where part of this work was carried out. We also thank Prof. Tukeyev for his valuable input. AS thanks the 2014 Google Summer of Code for her scholarship and mentor Inari Listenmaa for her assistance, and Aidana Karibaeva for help in writing transfer rules.

REFERENCES

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. Machine translation, 25(2):127–144.

Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila, A. (1995). Constraint Grammar: A language independent system for parsing unrestricted text. Mouton de Gruyter.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics–Doklady 10, 707–710. Translated from Doklady Akademii Nauk SSSR, pages 845–848.

Linden, K., Silfverberg, M., Axelson, E., Hardwick, S., and Pirinen, T. (2011). HFST–Framework for Compiling and Applying Morphologies, volume Vol. 100 of Communications in Computer and Information Science, pages 67–85.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sundetova, A., Forcada, M. L., Shormakova, A., and Aitkulova, A. (2013). Structural transfer rules for english-to-kazakh machine translation in the free/

[1] *https://svn.code.sf.net/p/apertium/svn/incubator/apertium-eng-kaz*

open-source platform apertium. In Proceedings of the International Conference on Computer processing of Turkic Languages, pages 317–326.

Sundetova, A., Karibayeva, A., and Tukeyev, U. (2014). Structural transfer rules for Kazakh-to-english machine translation in the free/open source platform Apertium. In Proceedings of the International Conference on Turkic language processing "TURKLANG'14", Istanbul Technical University, 6–7 November 2014, pages 91–96.

Tyers, F. M., Sánchez-Martínez, F., and Forcada, M. L. (2012). Flexible finite-state lexical selection for rule-based machine translation. In Proc. of the 16th Annual Conference of the EAMT, pages 213–220, Trento, Italy.

Washington, J., Salimzyanov, I., and Tyers, F. (2014). Finite-state morphological transducers for three Kypchak languages. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA).

Pecherskikh, T.F., Amangeldina, G.A.(2012). Features of translation of different languages (on example English and Kazakh languages). Young scientist, 3:259–261.

A.M. Sundetova and A. S. Karibaeva. (2013). Creating bilingual dictionary for English-Kazakh machine translation system on Apertium Platform. In Proceedings of the international scientific-practical conference "The application of information and communication technologies in education and science", dedicated to the 50th anniversary of the Department of Information and Communication Technologies and the 40th anniversary of the Department of "Information Systems" of Al-Farabi Kazakh National University. 22 november 2013, pages 53–57.

AUTOMATON MODELS OF THE MORPHOLOGY ANALYSIS AND THE COMPLETENESS OF THE ENDINGS OF THE KAZAKH LANGUAGE

Ualsher Tukeyev

Al-Farabi Kazakh National University
Almaty, Kazakhstan
ualsher.tukeyev@kaznu.kz

In this paper we investigate the question of the completeness of the endings in a source language in automaton models of morphological analysis of an agglutinative languages, in particular, the Kazakh language. A complete system of the endings of the Kazakh language is created. A comparative analysis with previously built the endings system of Kazakh by Bektayev is made. Completeness of endings of analyzed languages ensures that all the words of the input text will be correctly analyzed at the morphological level.

1. Введение

Вопрос морфологического анализа являются важными в процессе обработки естественных языков. Определяющим подходом морфологического анализа является двухуровневая морфология, предложенная (Koskenniemi, 1983), реализуемая через использование конечных преобразователей. Системы морфологического анализа языков, основанные на двухуровневой морфологии, используют аппарат теории автоматов, а именно, аппарат теории и методологии конечных преобразователей (finite state transducers-FST). Имеются публикации в области использования технологии двухуровневой морфологии и FST для агглютинативных языков (Oflazer, 1994; Washington et al., 2014; Kairakbay and Zaurbekov, 2013; Kessikbayeva and Cicekli, 2014). В работе (Тукеев и др., 2013) авторы использовали класс FST, а именно, табличные многозначные отображения для анализа морфологии инфлективных языков, относящиеся к недетерминированным FST.

В применении технологии FST для инфлективных языков существенным является определение множества возможных окончаний слов анализируемых языков. В случае построения FST в виде графового представления, множество возможных окончаний слов

будет представляться множеством возможных путей на графовой модели. В случае представления в виде табличных многозначных отображений, множество возможных окончаний слов будет представляться множеством строк таблицы многозначных отображений. В любом из этих представлении анализа морфологии весьма существенным является вопрос: является множество возможных окончаний слов анализируемого языка полным? Если, по тем или иным причинам, множество возможных окончаний слов анализируемого языка будет неполным, то возможна ситуация, что такое слово будет проанализировано неправильно или неуспешно. В связи с вышеизложенным проблема определения полноты множества возможных окончаний слов анализируемого языка является весьма важной.

Рассмотрим полноту множества окончаний на модели анализа морфологии казахского языка в виде следующих многозначных отображений:

$$F_s: X_s \rightarrow Y_s \text{ (для исходного языка)},$$
$$F_t: Y_t \rightarrow Z_t \text{ (для целевого языка)},$$

где X_s – окончания исходного языка,

Y_s – грамматические характеристики слова исходного языка,

Y_t – грамматические характеристики слова целевого языка,

Z_t – окончания целевого языка.

В данной системе отображений для обеспечения корректности преобразований любого слова пары языков машинного перевода необходимо, чтобы множество окончаний целевого и (или) исходного языка было полным. Полнота множества окончаний исходного языка весьма важным является для морфологического анализа предложений исходного языка, так как является гарантией того, что любое слово будет проанализировано на предмет его грамматической (лексической) характеристики.

В данной работе рассмотрим полноту системы окончаний казахского языка.

Так как полнота системы окончаний одного языка в паре машинного перевода определяет косвенно и в целом полноту системы преобразований на морфологическом уровне с одного языка на другой язык пары, то это является важным вопросом для всей системы машинного перевода.

2. Полнота системы окончаний казахского языка

Рассмотрим систему окончаний слов казахского языка двух классов: окончания к именным основам (существительные, прилагательные, числительные) и окончания к глагольным основам (глаголы, причастия, деепричастия, наклонения, залоги).

2.1. Система окончаний казахского языка к именным основам

Система окончаний к именным основам слов казахского языка имеет четыре типа:

– окончания множественного числа (обозначим через K),
– притяжательные окончания (обозначим через T),
– падежные окончания (обозначим через C),
– личные окончания (обозначим через J),
– основу(stem) обозначим через S.

Рассмотрим всевозможные варианты размещений типов окончаний: из одного типа, из двух типов, из трех типов и из четырех типов. Число размещений определяется формулой:

$$A_n^k = n!/(n-k)!.$$

Тогда, количество размещений будет определяться следующим образом:

$$A_4^1 = 4!/(4-1)!=4,$$
$$A_4^2 = 4!/(4-2)!=12,$$
$$A_4^3 = 4!/(4-3)!=24,$$
$$A_4^4 = 4!/(4-4)!=24.$$

Всего возможных размещений 64.

Рассмотрим какие из них семантически допустимы.

Размещения по одному типу окончания (K, T, C, J) являются все семантически допустимыми по определению.

Размещения по два типа окончаний могут быть следующие:

KT, TC, CJ, JK
KC, TJ, CT, JT
KJ, TK, CK, JC.

Анализ семантики размещений двух типов окончаний показывает, что выделенные жирным шрифтом размещения являются допустимыми (**KT, TC, CJ, KC, TJ, KJ**), а остальные размещения

относим к недопустимым. Например, ТК – после притяжательных окончаний окончания множественного числа не используются, СК – после падежных окончаний не принято ставить окончание множественного числа, JC – после личных окончаний не принято ставить падежные окончания, СТ – после падежных окончаний не ставятся притяжательные окончания, JT – после личных окончаний не ставятся притяжательные окончания. Отнесем к недопустимым и JK- после личных окончаний окончания множественного числа, так как этот тип размещения покрывается окончаниями множественного числа личных окончаний.

Вообще, типы окончаний **T** и **J** являются окончаниями определения зависимости субъектов, объектов, действий. В словах с именными основами для **TJ** двойное определение зависимости возможно для случаев различения субстанций (субъектов, объектов, действий): *апа-ң-мын (апа-ң относится к другому субъекту, а личное окончание –мын определяет зависимость к говорящему)*. В случае TJ двойное определение зависимости к одной и той субстанции запрещается, например: *апа-м-мын* не говорят.

Необходимо отметить, что тип окончаний **CJ** имеет ограничения по падежам ілік (родительный – genitive) и табыс (винительный – accusative).

Итак, количество допустимых (правильных) размещений из двух типов окончаний будет равно 6.

Размещения окончаний из трех типов будут следующие:

KTC, **KTJ**, **TCJ**, TCK, CJK, CJT, JKT, JKC

KCJ, KCT, TJK, TJC, CTK, CTJ, JTK, JTC

KJT, KJC, TKC, TKJ, CKT, CKJ, JCK, JCT.

Определение допустимых размещений окончаний из трех типов сделаем по правилу:

если в размещении из трех типов есть недопустимые размещения из двух типов, то это размещение – недопустимо.

Тогда, допустимых размещений окончаний из трех типов будет 4 (**KTC**, **KTJ**, **TCJ, KCJ** выделено жирным).

Размещения окончаний из четырех типов будут следующие:

KTJC, TKJC, CKTJ, JKTC

KTCJ, TKCJ, CKJT, JKCT

KJTC, TJKC, CTKJ, JTKC

KJCT, TJCK, CTJK, JTCK

KCTJ, TCJK, CJKT, JCKT
KCJT, TCKJ, CJTK, JCTK

Определение допустимых размещений окончаний из четырех типов сделаем по правилу:

если в размещении из четырех типов есть недопустимые размещения из двух типов, то это размещение – недопустимо.

Тогда, допустимых размещений окончаний из четырех типов будет 1 (**KTCJ** выделено жирным).

Итого, допустимых размещений из одного типа – 4, из двух типов – 6, из трех типов – 4, из четырех типов – 1.

Итак, суммарное число типов допустимых размещений в словах с именными основами – 15.

Ниже в таблице 1. представлены 15 типов окончаний слов казахского языка с именными основами с примерами и соответствующей грамматической структурой этих типов в русском языке с примерами. При описании соответствующей грамматической структуры на русском языке используются теги Penn Treebank (http://www.clips.ua.ac.be) и плюс PPRN – притяжательные местоимения, PPRNPL – притяжательные местоимения множественного числа.

Таблица 1

Типы окончаний слов казахского языка с именными основами с примерами и соответствующей грамматической структурой этих типов в русском языке

Типы окончаний казахского языка в словах с именными основами	Примеры на казахском языке	Адекватная грамматическая структура на русском языке	Примеры на русском языке
S-K	тәте-лер	N+PL	Тет-и
S-T	тәте –м	PPRN N	моя тетя
S-J	тәте –мін	PRN – N	Я-тетя
S-C	тәте –ге	IN N	К тете
S-K-T	тәте-лер-ім	PPRN N+PL	мои тети
S-K-J	тәте-лер-міз	PRN – N+PL	Мы-тети
S-K-C	Тәте-лер-ге	IN N+PL	К тетям
S-T-J	Тәте-м-сіз	PRN – PPRN N	Вы-моя тетя

S-T-C	Тәте-м-ге	IN PPRN N	К моей тете
S-J-K	Тәте-сің-дер	PRN – N	Вы-тети
S-C-J	Тәте-ден-сің	PRN IN N	Вы – от тети
S-K-T-J	Тәте-лер-ім-сіндер	PRN – PPRNPL N+PL	Вы – мои тети
S-K-T-C	Тәте-лер-ім-ге	IN PPRNPL N+PL	К моим тетям
S-K-C-J	Тәте-лер-ге-мін	PRN IN N+PL	Я к тетям
S-T-C-J	Тәте-ң-нен-біз	PRN – IN PPRN N	Мы- от твоей тети
S-K-T-C-J	Тәте-лер-ің-ге-міз	PRN – IN PPRNPL N+PL	Мы-к твоим тетям

2.2. Система окончаний казахского языка к глагольным основам

Система окончаний казахского языка к глагольным основам включает следующие виды:

– система окончаний глаголов;
– система окончаний причастий;
– система окончаний деепричастий;
– система окончаний наклонений;
– система окончаний залогов.

Система окончаний к глагольным основам (глаголы) включают следующие типы:

– времена (8 времен),
– лицо (3 вида),
– отрицание.

Тогда, количество возможных типов окончаний глаголов будет – 25.

Система окончаний к глагольным основам причастия включают следующие типы:

– окончания причастия (обозначим R),
– окончания множественного числа (обозначим К),
– окончания притяжательные (Т),
– падежные окончания (обозначим С)
– личные окончания (обозначим J).

Тогда, возможные варианты типов окончаний (тип окончаний причастия для всех вариантов одинаков) будут:

– с одним типом окончаний:

RK, RT, RC, RJ;

– с двумя типами окончаний:

RKT, RTC, RCJ, RJK
RKC, RTJ**,** R*CT, RJT*
RKJ, R*TK, RCK, RJC;*

– с тремя типами окончаний:

RKTC, RTCJ, RCJK, RJKT
RKTJ, RTCK, RCJT, RJKC
RKCJ, RTJK, RCTK, RJTK
RKCT, RTJC, RCTJ, RJTC
RKJT, RTKC, RCKT, RJCK
RKJC, RTKJ, RCKJ, RJCT;

– с четыремя типами окончаний:

RKTJC, RTKJC, RCKTJ, RJKTC
RKTCJ, RTKCJ, RCKJT, RJKCT
RKJTC, RTJKC, RCTKJ, RJTKC
RKJCT, RTJCK, RCTJK, RJTCK
RKCTJ, RTCJK, RCJKT, RJCKT
RKCJT, RTCKJ, RCJTK, RJCTK.

Рассмотрим семантическую допустимость вариантов окончаний.

Все варианты окончаний причастий по одному типу окончаний являются семантически допустимыми.

Анализ семантики размещений двух типов окончаний причастий показывает, что выделенные жирным шрифтом размещения являются допустимыми, а остальные размещения относим к недопустимым. Допустимые варианты окончаний причастий такие же, как в системе окончаний с именными основами, но из них для причастий являются недопустимым вариант RTJ, так как последовательность «окончание причастия-притяжательные окончания» для причастий во всех случаях означает персонофицированное действие с глагольной основой. А персонофицированное действие не может второй раз персонофицироваться личным окончанием. Например, *бар-ған-ым(мой приход, ту coming)*, однако нельзя сказать *бар-ған-ым-сың(ты – мой приход)*, так как действие *(бар-ған-*

ым) не персонофицируется, т.е. действие не может представиться субъектом.

Аналогично, окончания RTCJ и RKTCJ не имеют ограничений по двум типам окончаний, т.е. возможные пары окончаний внутри этих типов окончаний являются допустимыми, но они нарушают предыдущее правило «действие не может представиться субъектом». Например, для RTCJ: *бар-ған-ым-ға-мын*, где «*бар-ған-ым-ға»(к моему приходу – to my coming)* – склонение действия, что не может представиться субъектом. Для RKTCJ: *бар-ған-дар-ың-нан-біз*, где «*бар-ған-дар-ың-нан*»(*от твоих приходов – from yours coming*) – склонение действий, что не может представиться субъектами.

Таким образом, количество типов окончаний причастий составляет – 11.

Рассмотрим типы окончаний деепричастий. Они представляются окончаниями переходного будущего времени, за которыми следует личные окончания: PJ, где P – базовое окончание деепричастия, J – личные окончания. Для данного класса выделим только следующие базовые окончания: *-ғаны, -гелі, -қалы, -келі.* Таким образом, считаем, что количество типов окончаний деепричастия -1.

Рассмотрим окончания наклонений, а именно, условного, повелительного, желательного. Окончания изъявительного наклонения совпадают с окончаниями глаголов в настоящем, прошлом и будущем.

Тип окончаний склонений аналогичен предыдущему, т.е. базовые окончания наклонений, за которыми следуют личные окончания. Таким образом, будем считать, что имеются три типа окончаний наклонений: условного, повелительного, желательного.

Типы окончаний залогов, а именно, возвратного, страдательного,совместного и принудительного, также определяются по предыдущей схеме: базовые окончания залогов за которыми следуют личные окончания. Соответсвенно, типов окончаний залогов будет – 4.

Итак, общее количество типов окончаний слов с глагольными основами будет 48.

Итого, общее количество окончаний с именными основами плюс общее количество типов окончаний слов с глагольными основами будет равно 63.

Следующей задачей является по полученным типам окончаний определить формы окончаний и их количество. Это сделать несложно так как для каждого типа части речи имеются соответствующие правила. В данном направлении автором построены конечные множества окончаний для всех основных частей речи казахского языка. Так, для частей речи с именными основами количество окончаний равно 862, а количество окончаний частей речи с глагольными основами составляет: глаголы – 432, причастия - 1588, деепричастия - 48, наклонения – 230, залоги - 80. Итого, 3240 всего окончаний.

3. Сравнительный анализ разработанной системы окончаний казахского языка с моделью Бектаева

Данная работа является развитием модели Бектаева (Бектаев, 1999) для применения в области машинного перевода. В модели Бектаева определено множество окончаний казахского языка в количестве 753 окончаний. Бектаев в своей модели предложил также алгоритм использования разработанной им системы окончаний казахского языка в связи их с грамматическими характеристиками для правильного и точного перевода (немашинного) на другой язык. В модели Бектаева определены 15 типов окончаний слов с именными основами. Предлагаемая модель также имеет 15 типов окончаний, однако отличается двумя типами: в модели Бектаева используются типы JK и TJK, которые нами отнесены к окончаниями множественного числа личных окончаний. В модели Бектаева для слов с глагольными основами используются три типа окончаний с причастиями, в то время как в предлагаемой модели – 11 типа окончаний с причастиями, один тип деепричастия, три типа окончаний наклонений: условного, повелительного, желательного, четыре типа окончаний залогов, а именно, возвратного, страдательного, совместного и принудительного. В общем итоге, количество окончаний в предлагаемой модели 3240 против 753 окончаний модели Бектаева.

4. Заключение и дальнейшие работы

В работе сделана попытка построения полной системы окончаний казахского языка, что будет являться основанием для полноты

системы морфологического анализа текстов на казахском языке. Так как казахский язык относится к агглютинативной группе языков, то вопрос полноты системы окончаний других языков данной группы может быть исследован аналогично. Данный подход, по мнению автора, позволит повысить качество морфологического анализа, соответственно, и качество машинного перевода. В качестве дальнейших работ планируется использовать результаты данного исследования в разработке систем машинного перевода, проводимых автором и его исследовательской группой.

Данные исследования проводятся в рамках грантового финансирования 0749/ГФ4 Министерства образования и науки Республики Казахстан.

REFERENCES

Koskenniemi, K. (1983). *Two-level morphology: A general computational model of word-form recognition and production*. Tech. rep. Publication No. 11. Department of General Linguistics. University of Helsinki.

Gurenko, V.V. (2013). *Intoduction to automata theory*. Electronic handbook. – M.: MGTU, -pp. 62 (на русском языке).

Oflazer, K. (1994). *Two-level description of Turkish morphology*, Literary and Linguistic Computing Volume9, Issue2. 137–148.

Washington, J. N., Salimzyanov, I., Tyers, F.M. (2014). *Finite-state morphological transducers for three Kypchak languages.* Proceedings of the 9th Conference on Language Resources and Evaluation.

Kairakbay, B.M., Zaurbekov, D. L. (2013). Finite *State Approach to the Kazakh Nominal Paradigm.*. Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing. St Andrews–Sctotland. 108–112.

Kessikbayeva, G., Cicekli, I. (2014). Rule *Based Morphological Analyzer of Kazakh Language*. Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM, Baltimore, Maryland USA. 46–54.

Тукеев, У.А., Рахимова, Д.Р., Байсылбаева, К., Умирбеков, Н., Оразов, Б., Абақан, М., Кызырканова, С., (2013). Көпмағыналық бейнелеу кесте тәсілі негізінде орыс тілінен қазақ тіліне машиналық аудармасының морфологиялық анализбен синтезін құру. Түркі тілдерін компьютерлік өңдеу. Бірінші халықаралық конференция: Еңбектері / Астана: Л.Н.Гумилев атындағы ЕҰУ баспасы, 182–191.

Bektayev, K. (1999). *Big Kazakh-Russian and Russian-Kazakh dictionary*, Almaty, “Altyn Kazyna”. pp.704 (in Kazakh, Russian).

SECTION 2 | **LINGUISTIC SOFTWARE**

STUDY ON FREQUENCY STATISTIC OF KAZAKH COMMON-USE WORD

Gulila Altenbek

College of Information Science and Engineering,
Xinjiang University, Urumqi, Xinjiang, 830046, China
Xinjiang Laboratory of Multi-language Information Technology
The Base of Kazakh and Kirghiz Language
of National Language Resource Monitoring
and Research Centre Minority Languages,
gla@xju.edu.cn

Kazakh language is an agglutinative language with word structures formed by productive affixation of derivational and inflectional suffixes to stems. With automatic extraction of Kazakh common-use words, we used a special formula to calculate the frequency of Kazakh common-use words. Experimental results show that the improved calculation formula has greater ability to rank word position than the traditional method used to extract Kazakh common-use words.

1. Introduction

Frequency statistic and extraction of word are important tasks in Kazakh natural language processing research. Morphological analyzers have been developed for different languages, which includes the

Porter Stemmer for English (Porter,1980), the PC Kimmo for Finnish (Karttunen ,1983), Hankamer's Keci for Turkish(1986), Beesley for Arabic(1996).In a national language system, vocabulary is the basic carrier of language information and the most active, vital elements in the language system (Zhao Xiao-Bing, 2007). Common-use words are those used most frequently in everyday language and are a relative and stable open set.

Kazakh Language belongs to the Turkish Language group in the Altaic language family, and it is an agglutinative language with word structures formed by adding derivational or inflectional affixes to root words. The Kazakh nationality are spoken same Kazakh language as a native language all around the world, but there are three different writings of Kazakh characters: Cyrillic letter is used as national language for Kazakh alphabet in Kazakhstan; native speakers in P. R. China use Arabic letter for Kazakh alphabet, and in Turkey, Latin letter is widely used. Previous work done on Kazakh language processing research for Kazakh Arabic letter (Altenbek et al.2009–2010, 2014.). Other have design and compilation process of Corpus (Makhambetov et al.2013, Doszhan et al. 2013) and Finite-state morphological transducers (Washington Washington et al.2014) for Kazakh Cyrillic letter. According to our knowledge, there is no research on Suffix-based Part of Speech Tagger for Arabic letter Kazakh. This paper is a first time to do research work for Arabic Kazakh POS tagger based on Morphological feature.

In this paper, the research mainly focuses on Kazakh common-use words based on Morphological feature, which are the most difficult aspect of Kazakh natural language processing using the statistical method by Arabic script in P. R. China. we have using three features of common-use words, that is, filed generality, time generality and regional generality, which is based on the corpus from mainstream newspaper media and Kazakh school Textbooks. The Kazakh have a rich vocabulary, and the research of Kazakh common-use word conforms to the social development. To make statistical analysis of a lot of different styles of written and spoken materials, and to use computer technology to study the Kazakh vocabulary in scientific research, it is important to identify the Kazakh common-use words.

2. Kazakh common-use word and its characteristics

2.1. Kazakh common-use word

Common-use words are Kazakh people's daily use, not easy to change, and relatively stable, it is the core of language. They have a close relationship to the daily life of the people and carry over between generations. They include terms for natural phenomenon, livestock, family name, parts of the human body, seasonal, activities number, work tools and so on. The Kazakh people have been engaged in livestock husbandry, and have had a nomadic life, so the Kazakh common-use words also contained a large vocabulary about animal husbandry.

The definition of Kazakh common-use words is the following in our paper: the using word of Kazakh in daily communication, mainly involving the words which have high frequency usage and high degree in each field, each local interval and each period.

2.2. Kazakh common-use word's characteristics

According to the definition of common-use word, we can get the common-use word's characteristic of generality, as the following :

1) Field generality: Common-use word has the character of widespread use in all fields and trades;

2) Regional generality: Common-use word has the character of communicating wide spread with the people of different regions;

3) Time generality: Common-use word has the character of time stability.

Based on the three characteristics, we use the 'lexical general usage' indicators to quantize the description of these characteristics. Using the statistical method to investigate the general level of Kazakh universal vocabulary, which used by the mass media, in order to realize the goal of common-use word's automatic extraction for the Kazakh mainstream newspaper media based on the corpus.

3. Kazakh Morphology

3.1. Kazakh Morphology

Kazakh Words morphological changes consist by large number of suffix and a small amount of prefix. Every word has a root, or a

stem (Milat , 2003). Kazakh word can be divided into stems and suffixes.

E.g. word= prefix+ stem +affixes.

The root is the core of the entire word structure and it conveys its basic meaning as a single monolithic morpheme.

A stem is a new word generated by adding one or more than one various affixes to the root, and it expresses the complete meaning of the word.

Affixes are divided into inflectional affixes and derivational affixes leading to a very large vocabulary size. Inflectional affixes, when added to a word do just cause grammatical changes, but does not lead to changes in meaning. Derivational affixes changes the meaning of the word when added to a root word.

Prefix is affix added before to the root or stem, it has only two and from loanwords that meaning negative by Kazakh language.

3.2. Kazakh orthography

The Kazakh language is written from right-to-left in the Arabic script in P.R. China. There are 35 sounds and 33 letters in this alphabet, including 9 vowels and 24 consonants. There are strict rules governing the order in which morphemes are combined.One syllable in Kazakh language is composed of a vowel surrounded by several consonants. However, one vowel can be a syllable. This means that the number of vowel equal to the number of syllable.

The basic forms of Kazakh syllables are the following: (1)A (ا), (2) AB(ات), (3)BA (جانه), (4) BAB (باس), (5) ABB (ایت), (6) BABB (کسلت). In the case of loanwords of Kazakh language, there are other forms of Kazakh syllables, which are BBA (پرولەتاریات), BBAB (تراکتور), BBABB (ترانسکریپسیا), BBBAB (ستره‌لکا) and so on. In this case, ‘A’ represents a vowel while ‘B’ represents a consonant.

4. Automatic extraction of Kazakh common-use words

Kazakh common-use words have three characteristics: filed generality, time generality and regional generality, and use these characteristics corresponding to the field general usage, time general usage and regional general usage of quantitative indicators to measure the degree of general vocabulary.

4.1. Quantitative analysis method of field general usage

Field general usage of words is used to measure the general level in each field of circulation of language, which is the quantitative index of commonly used words degree. The computation formula not only should examine the frequency of a word, but also should consider whether words in different areas, different texts and the distribution of the fields are uniform. Mainly include two kinds of qualitative and quantitative investigation ways:

1) Qualitative investigation: obtain all the fields of common words relying on fields' intersecting words.

2) Quantitative investigation: calculate the word's field general usage according to frequency and distributed or other situation in each domain.

Qualitative investigation ways are relatively simple, for the field distribution of vocabulary is also clear at a glance, and its shortcoming is ignoring the words frequency–an important indicator to measure the common degree of words, so this kind of investigation ways only as the assistant to provide reference. We will use quantitative calculation ways to analyze the field general usage of words.

4.1.1. The traditional calculation method of Kazakh words' field usage

In order to measure the degree of commonly used words, the traditional method is used in addition to calculate the use frequency of words, even integrated considerate the calculation words with different areas of the text and area. The calculation procedure is as follows:(1) Calculate word frequency F_k : F_k stands for total appearing frequency in the corpus of word k.

(2) Calculate the usage U_k of word k: Use the improved calculating format of Chang Baoru's word usage:

$$S_k = \sqrt{\sum_{i=1}^{n} (N_k^i - N_k)^2 / n}$$

$$D_k = 1 - S_k / (N_k \times (n-1)^{\frac{1}{2}}) \quad (0 \le D_k \le 1)$$

$$DI_k = \frac{P_k + L_k \times C_1 + C_2}{P + n \times C_1 + C_2} \text{ (Word comparative frequency<0.0001)}$$

$$DE_k = \frac{1}{2} DI_k + \frac{1}{2} D_k \text{ (Word comparative frequency≥0.0001)}$$

Calculating formula of words fields usage: Words usage U_k=DE_k (or DI_k)× F_k (take integer). Note: D_k stands for the spread coefficient of word k. DE_k stands for the high frequency band of spread coefficient of word k. DI_k stands for the low frequency band of spread coefficient of word k. N_k^i stands for the relative frequency of word k in the field of I. Nk stands for the total relative frequency of word k in all fields. P_k stands for distribution number of articles of word k. C_1 and C_2 are pending constant. P stands for the total articles of corpus. N stands for classification number field of corpus. Obviously, $1 \le P_k \le$P, $1 \le L_k \le$n. Make n× C_1 =P, C_1 can be sure; Make $0.5 \le DI_k \le 1$, C_2 can be sure, C_1 and C_2 take a positive number.

Because the goal of this paper is to calculate the usage of each month, so P value is to process the total text of every month. By the validation of Xinjiang daily of Kazakh edition in 2008, this paper set s values as C_1 =P/n, $C_2 \approx 2 \times$P, make C_2 =1200.

The above formula is considering the word frequency, word text spread and word field distribution and other integrated circumstance to obtain quantized index of commonly used features of words---word usage, this is consistent with the goal of the words commonly used degree, so it can be used as the calculation formula of lexical general usage.

The traditional method has the following aspects of defects:

1) Fixed word text distribution and field distribution together, it is needed to calculate two constants of value C_1 and C_2, and their values have a relationship to the total number of corpus and the domain classification number, so the formula is more suitable for application in a closed static corpus. Regarding the corpus which is used in this experiment, the total number of the corpus and domain classification numbers are always in dynamic status updates, so should not use the formula.

2) Even in the total number of the corpus and the domain classification number under certain conditions, the value of c1 can be uniquely identified, but the value of c2 is required for a range of values, so the value of c2 is not only, it also gives the calculation results bring uncertainty.

3) Even though the field distribution characteristics in this formula have responses, their performance on the word general extent is relatively weak. But the lexical general usage requires that the field distribution parameters have more influence on measuring general degree of words.

4.1.2. The improved calculation method of Kazakh words' field general usage

In order to solve the weakness of traditional calculation formula of filed general usage, improve the calculation method of filed general usage. The calculation steps of improved filed general usage are as follows:

1) Compute word frequency of field: F_k stands for the total frequency of field classification corpus of word k,

2) Calculate the usage of word K in text: Use A. Juilland's formula to calculate the usage of words in text categorization:

$$S_k = \sqrt{\sum_{i=1}^{n} (N_k^i - N_k)^2 / n}$$

$$D_k = 1 - S_k / (N_k \times (n-1)^{\frac{1}{2}}) \quad (0 \le D_k \le 1)$$

The usage of words $UL_k = D_k \times F_k$ (take integer values).

Note: N_k^i stands for the relative frequency of word k in the field I. Nk stands for the total relative frequency of word k in all fields. N stands for the total text number of corpus. D_k stands for the coefficient of dispersion of word k. F_k stands for the frequency of word k.

3) Calculate the field general usage of word U_k : Use the Distributional Consistency (DC) to calculate the even degree of word in all areas. Calculation formula is that the Distributional Consistency DC_k=SMR/Mean ($0 \le DC_k \le 1$)

The definitions of SMR and Mean are as follows:

$$SMR = (\sum_{i=1}^{n} \sqrt{FK_i} / n)^2 \qquad Mean = (\sum_{i=1}^{n} Fk_i / n)$$

The field general usage of words k $^{U_k} = DC_k \times UL_k$.

Note: n stands for the number of field classification, and requires that the number of field of corpus is equal, FK_i stands for the frequency of word k in the field I. UL_k stands for the text usage of words k. DC_k stands for the even degree of word k in the field classification.

4.2. Calculation method of time general usage

The time general usage is the quantitative of words in the investigation time. It needs to observe whether the words in the inspection period are stable or not, namely the even degree of words in every month.

The procedure of time general usage is as follows:

1) Calculate the monthly frequency of words: F_k stands for total appearing frequency in every month for word k.

2) Calculate the time general usage of word k: Use Distributional Consistency (DC) to calculate equal degree of words in every month of the distribution in the investigation, the calculation formula is:

$$SMR = (\sum_{i=1}^{n} \sqrt{FK_i} / n)^2 \quad Mean = (\sum_{i=1}^{n} Fk_i / n)$$

The time general usage of word k T_k = SMR/Mean ($0 \leq T_k \leq 1$)

Note: n stands for the number of months in investigation, the requirement of each month's corpus is equal; FK_i stands for the frequency in the month i.

4.3. Feature description of region general usage

Region general usage of word is the vocabulary's use situation in different regions of the media from the point of view of common time, that is, during the investigating time, the stability degree of vocabulary in different regions medial.

Region general usage is similar to the time general usage. They are also investigating the equality degree in different distribution of the classification system, which is the stable degree of using except that the difference is that time general usage is classified by months, but region general usage is classified by the different parts of the media.

4.4. Calculation method of lexical general usage

As stated above, the description of the 'lexical general usage O_k' is considering the field general usage and time stability of words, and but does not take into account the influence of region general usage to lexical general usage.

Calculation formula of lexical general usage is:

Lexical general usage $O_k = T_k \times U_k$.

Note: T_k stands for the time general usage of the word k. U_k stands for filed general usage of the word k. O_k stands for general level of

word, the higher of the O_k value, the characteristic of common use and the stability feature of inspected time using performance are better.

5. Extract process of Kazakh common-use words and Implement of automatic extraction system

5.1. Data of Corpus

In the experiment, we used the data of 2008 year of *Xinjiang daily* (Kazakh version) and Kazakh Textbooks from the Xinjiang University Kazakh corpus. The corpus consists of raw texts and the TXT or XML format texts. In order to study the vocabulary field general, the original media corpus classification is needed. This paper will divide the raw corpus into 5 classes as political, economy, education, life and sports. The extraction of Kazakh common-use words mainly divides into two modules: preprocessing module and extraction module.

5.2. System implementation

According to the lexical general usage calculation method of Kazakh vocabulary, use C# language for system development, and recognize the Kazakh lexical general usage statistical system. Calculation process of Kazakh common-use words and system interfaces is as follows:

Fig.1. Kazakh word frequency statistic system

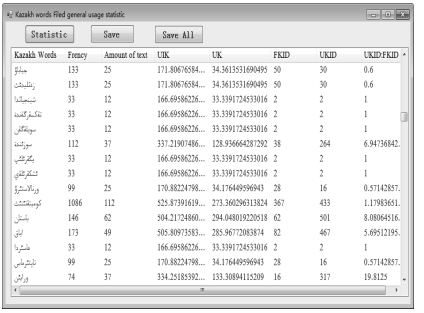


Fig.2. Kazakh word field general usage statistic system

1) Calculation of fields, each Kazakh text ‘word’ of the word frequency, figure 1 shows the system interface of Kazakh word frequency statistic.

2) Calculate text usage and field general usage in all areas' words monthly. Figure 2 shows the system interface of Kazakh word field general usage statistic.

3) Calculation of annual Kazakh word time general usage and lexical general usage. Figure 3 shows the system interface of Kazakh word lexical general usage statistic.

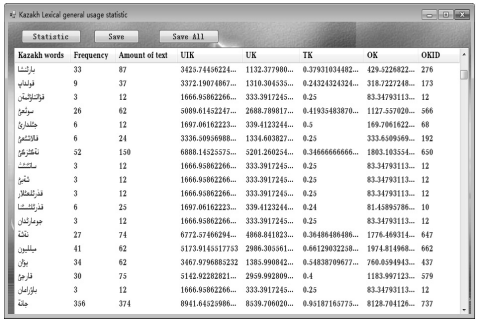


Fig.3. Kazakh word time general usage and lexical general usage statist system

6. Experiment results and analysis

According to the lexical general usage calculation method of Kazakh vocabulary, use C# language for system development, realize the Kazakh lexical general usage statistical system. In accordance with the Kazakh common-use words extraction rules, successfully extract Kazakh common-use words.

6.1. The analysis of improved field general usage calculation method

As mentioned previously, calculation formula of field general usage can use the traditional calculation formula of word usage, but we discussed this formula's defects, especially the impact of the words' field distribution on field level of general quantitative calculation is not obvious, so this study proposed an improved formula.

In order to validate improved calculation formula, we select the Xinjiang daily of Kazakh edition in January 2008 as the corpus of test data, separately calculate the field general usage of traditional formula and improved formula, and make comparing analysis of this test.

Compared experimental analysis of frequency:

This paper's frequency is the unified tag of FKID and UKID, namely the corresponding in the words table according to sort of position after order of 'the frequency of words' and 'field general usage'.

UKID: FKID > 1, it is shown that the position of the words compare to the original FKID position is on the backward adjustment, the higher the value, the greater the adjustment of the word backward, then the text spread and the stability of the field distribution usage for the word is not good.

UKID: FKID < 1, it is shown that the position of the words compare to the original FKID position is on the forward adjustment, the higher the value, the greater the adjustment of the word forward, then the text spread and the stability of the field distribution usage for the word is good.

UKID: FKID = 1, it is shown that the position of this words is same as the original position of FKID, there is no effect for the range position of this word. The calculation formula of traditional field usage.

Table 1

The part results of backward words of the traditional calculation formula of field usage

Kazakh words	FK	Text number	UK	FKID	UKID	UKID:FKID
بولدى	494	136	527.895608	276	630	2.282609
باستاماشى	246	50	129.506097	138	268	1.942029
قوعام	391	100	288.097844	231	477	2.064935
وسى	1986	361	827.966280	390	668	1.712821
حالىق	625	174	800.304715	311	659	2.118971
كوركەمونەر	240	50	119.456207	134	214	1.597015
جوق	525	149	518.871667	285	616	2.161404

Table 1 is the processing experimental results of the traditional calculation formula of field usage, cutting part of backward words.

Table 2

The part results of forward words of the traditional calculation formula of field usage

Kazakh words	FK	Text number	UK	FKID	UKID	UKID:FKID
شارۋالار	368	62	109.049662	221	173	0.782805
باعاسى	232	25	34.756044	129	62	0.480620
بيگه	199	25	34.778509	104	64	0.615385
قىستاقتاردىڭ	335	62	110.70047	202	176	0.871287
ورىندادى	99	37	34.293415	28	24	0.857143
بارىنشا	340	87	113.13839	205	184	0.897561
تەڭشه	99	25	34.176450	28	16	0.571429

Table 2 is the processing experimental results of the traditional calculation formula of field usage, cutting part of the adjustment forward Kazakh words.

The maximum value of UKID: FKID is 2.282609, the minimum value is 0.480626, so adjustment range of words' position for traditional computational formula is not big. Thus the text spread and field distribution does not bring about too much impact to sort result of words.

Table 3

The part results of backward words of the calculation formula of improved field general usage

Kazak word	Text number	FK	FKID	UK	UKID	UKID:FKID
دۇنيه	424	4957	1	859.3169442	5	5
شارۋا	374	3587	2	855.0248183	6	3
14	175	1905	9	137.4895981	322	35.77778
بەيلىق	265	1850	11	288.3072141	302	27.45455
مۇناي	162	1973	12	457.1887243	106	8.833333
25	112	1368	22	142.8154444	309	14.04545
ساۋدا	87	1036	37	105.8061292	889	24.02703
ولىميا	50	887	50	35.98305303	807	16.14

Table 3 is the processing experimental results of the calculation formula of improved field general usage, cutting part of backward words.

Table 4

The part results of forward words of the calculation formula of improved field general usage

Kazakh words	Text number	FK	FKID	UK	UKID	UKID:FKID
بۇل	386	2932	4	871.8128425	1	0.25
اۇىل	49	113	985	300.4290849	142	0.144162
ەارى	100	336	215	533.8430249	39	0.181395
ەل	124	353	299	762.0626994	25	0.083612
مەدەنيەت	74	237	333	504.7022196	79	0.237237
جۇگگو	99	264	349	477.628399	94	0.269341
دەپ	87	256	361	520.531814	57	0.157895
ۇرىمجى	98	234	396	517.3909156	67	0.169192
شينجياڭ	74	232	405	493.3513744	86	0.212346
قازاق	74	288	425	517.3553599	68	0.16

Table 4 is the processing experimental results of the calculation formula of improved field general usage, cutting part of the adjustment forward Kazakh words.

The maximum value of UKID:FKID is 35.77778, the minimum value is 0.083612, the adjustment range expands, so adjustment range of words' position for improved calculation formula is larger. Thus the impact strength of the text spread and field distribution of words is equal or not for the adjustment position of word and will increase.

The experimental results show that the traditional calculation formula does not change the adjustment position of word, therefore, the text spread and field distribution of words do not bring about too much impact on sort result of words. However, improved calculation formula changes the adjustment range of position of word a lot, thus the impact strength of the text spread and field distribution of words is equal or not for the adjustment position of word will increase, this is more accord with our recognition of the common-use words.

From the experimental results, the extraction method is basically satisfactory, but the accuracy of lexical general usage still remains at rising further. In this experiment, the original corpus is divided into 5 categories: political, economy, education, life, sports.

6.2. The analysis of word statistics Based on corpus

Kazakh language has a complex morphological structure, similar to typical agglutinative languages. Words can be formed by long concatenations of morphemes with some order or semantic features.

This study has been completed statistical content: kazakh words starting with statistics, the kazak of word frequency statistics, the kazakh word length statistics, all the relationship between the length and frequency of the word. The experimental results illustrate the kazak inner link between word frequency, and at the same time to verify the kazakh word frequency comply with Zipf's law of power law.

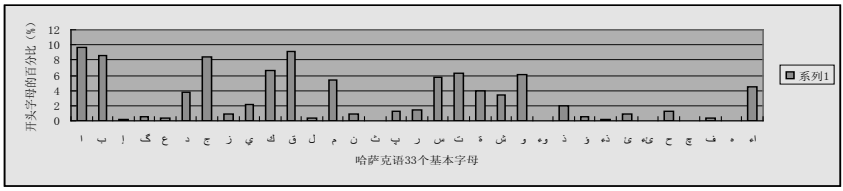


Fig.4. Initials letter statistics of Kazakh words

Figure 4 show the letter ا accounts for 9.71% of the total vocabulary of the courps, then accounted the letter ق and ب for 9.09% and 8.66%, the letter ھ is minimum number of words Accounts for only 0.019% of the total vocabulary only in interjection.

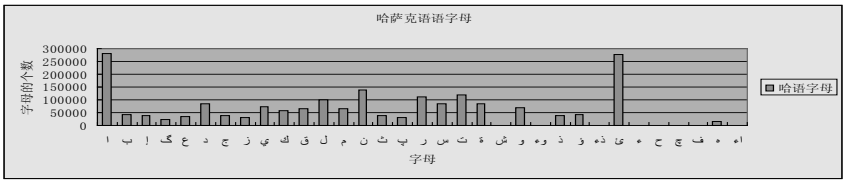


Fig.5. All letter statistics of Kazakh words

Figure 5 show the letter ا is 28.2149 ten thousand letter, ى is 27.8275 ten thousand letter and then ن is 14.7267 ten thousand letter. ھ is only78,is 104.

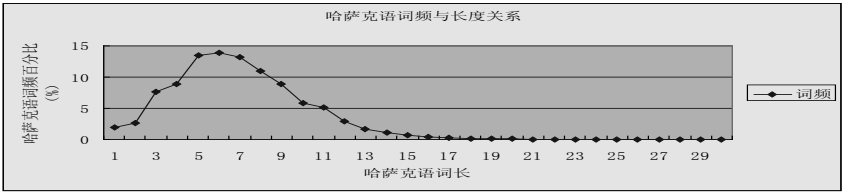


Fig.6. The relationship between the length and word frequency

Table 5

The relationship between the words of length and frequency

length / frequency (%)	1–4	5–8	9–12	13–30
The corpus words total statistics is 4.62MB	1–2.0054	5–13.4878	9–8.9305	13–1.6849
	2–2.8539	6–13.8508	10–5.7958	23–0.0355
	3–7.5870	7–13.1275	11–5.0928	26–0.0257
	4–8.8262	8–10.9275	12–2.8579	30–0.0146
total	21.0025	51.3936	22.6770	4.9289

Table 6

Part of most high frequency words from textbooks

Middle school			High school		
order	word	frequency	order	word	frequency
1	ری ب ُ	2 324	1	ری ب ُ	3524
2	اد	1 949	2	اد	2788
3	پ ەد	1 724	3	ن ەم	2446
4	ن ەم	1 628	4	پ ەد	2329
5	ەد	1 494	5	ەد	2183

6	ولا	1 225	6	لؤلب	1520
7	هدی	920	7	ولا	1491
8	لؤلب	917	8	بولادی	1197
9	هكەن	853	9	دەگەن	1193
10	وسی	832	10	راب	1183

Table 7

Part of most high frequency stems from textbooks

Middle school			High school		
order	word	frequency	order	word	frequency
1	دە	3774	1	دە	5181
2	لوب	2878	2	لوب	4381
3	ریبُ	2674	3	ریبُ	4325
4	اد	1953	4	اد	2830
5	نەم	1918	5	نەم	2674
6	راب	1669	6	راب	2386
7	زوُ	1376	7	زوُ	2054
8	ولا	1371	8	رؤت	1752
9	هكەن	1217	9	ولا	1702
10	بەر	1166	10	بەر	1605

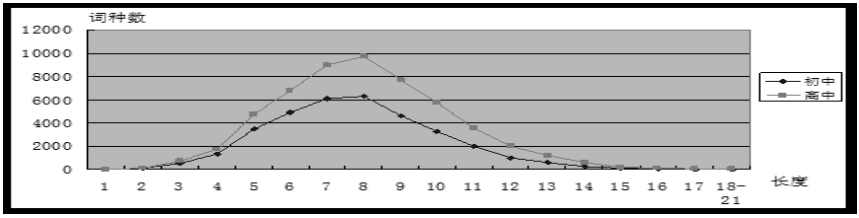


Fig.7. The relationship between the word of length and frequency for Textbooks

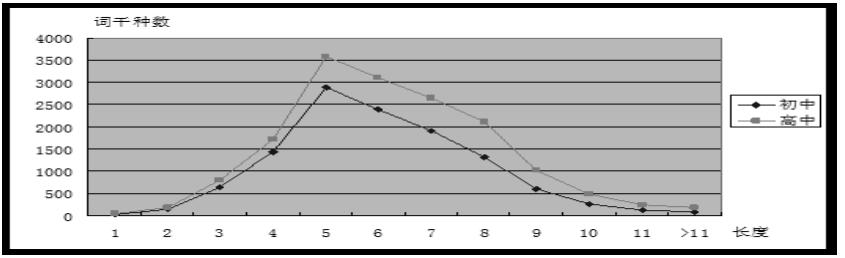


Fig.8. The relationship between the stem of length and frequency for Textbooks

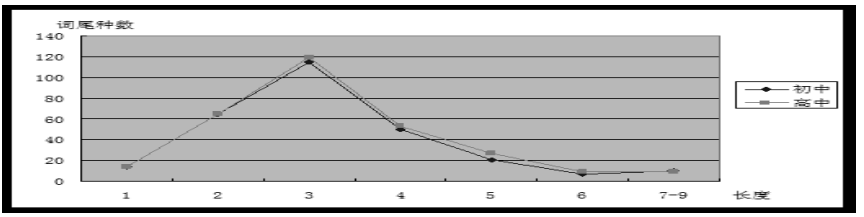


Fig.9. The relationship between the affix of length and frequency for Textbooks

The relationship between the length and frequency data from textbooks corpus in Figure 7–9, other figure and table date from Xinjiang Daily. As about that the relationship between length and frequency of all word has a remarkable characteristic is to the left, which means most of the word length is short, E.g. On the one hand, The length from 5 to 8 accounting is 51.3936% of all words, from 13 to 30 accounting is only 4.9289% of all words, on the other hand, Dragging a long "tail" is the another big characteristic of power law.

7. Conclusion

This article used the definition of the Kazakh common-use words, and identified the Kazakh common-use words' three characteristics: filed general**ity** ,time general**ity** and regional general**ity** , and used these characteristics corresponding to the field general usage, time general usage and regional general usage of quantitative indicators to measure the degree of general vocabulary. Based on the improved words filed general usage to achieve a Kazakh common-use words extraction, suc-

cessfully extract Kazakh general vocabulary, and has good extraction effect, make up for the traditional words filed usage is insufficient, make words' field distribution characteristics have more influence of the words' commonly used characteristics.

In the experiment, we used the data of 2008 year of Xinjiang daily (Kazakh version) and Kazakh Textbooks from the Xinjiang University Kazakh corpus. On the basis of frequency statistics of Kazakh words from Kazakh corpus, we derived a formula for Kazakh lexical general usage. Experimental results show that the improved calculation formula has greater ability to rank word position than the traditional method used to extract Kazakh Common-use words, also the result expresses the relation of frequency of the Kazak word, and the resulting Kazakh word frequency distribution accords with power-law of Zapf.

Acknowledgements

This work is funded by the Natural Science Foundation of P.R. China (NSFC)(No.61363062, No. 61063025)

REFERENCES

Porter,M.F.1980. An algorithm for suffix stripping, Program, 14(3):130−137.

Karttunen,Lauri.1983. KIMMO: A general morphological processor. Texas Linguistic Forum, 22:163–186.

J.Hankamer, 1986. Finite state morphology and left to right phonology. In Proceeding of Fifth west coast conference on formal linguistics, 29–34.

Beesley, K.R. 1996. Arabic finite-state morpholog-ical analysis and generation. In COLING-96, Co-penhagen, 89–94.

Kemal Oflazer.1994. Two-level description of Turkish morphology. Literary and Linguistic Computing, 9(2):137–148.

Gülşen Eryiğit and Eşref Adalı, A.(2004).An affix stripping morphological analyzer for Turkish, Pro-ceedings of the International Conference on Artificial Intelligence and Application, Austria, 299–304.

Milat, A. 2003. Modern Kazakh language, Xinjiang People's press, China.

Gulila Altenbek. and Dawel,A. and Muheyat,N. 2009. A Study of Word Tagging Corpus for the Modern Kazakh Language, Journal of Xinjiang University, 26(4):394–401.

Gulila Altenbek, Ruina-Sun. 2010. Kazakh Noun Phrase Extraction based on N-gram and Rules, International Conference on Asian Language Processing (IALP2010). Harbin, China. 305–308.

Gulila Altenbek, Xiaolong Wang and Gulizha-daHaisha. 2014. Identification of Basic Phrases for Kazakh Language using Maximum Entropy Model. In: Proceedings of the 25th International Conference on Computational Linguistics (COLING), 23–29 August 2014, Dublin, Ireland. 1007–1014.

Olzhas Makhambetov, Aibek Makazhanov, Zhan-dos Yessenbayev,Bakhyt Matkarimov, Islam Sabyrgaliyev, and Anuar Sharafudinov. 2013. Assembling the Kazakh Language Corpus, In Proceedings of the Conference on Empirical Methods in Natural Language Processing 2013 (EMNLP 2013). Association for Computational Linguistics,1022–1031.

Gulzhan Doszhan,2013. Problems of Creation of the All-Turkic National Corpus. International Conference on Information, Business and Education Technology.1018–1023.

Jonathan North Washington, Ilnar Salimzyanov and Francis M. Tyers. 2014, Finite-state morphological transducers for Kypchark language, In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 3378–3385.

Qiangjun Wang, Isabella Park and Pu Zhang. 2003. Automatic Extraction of the Unlisted Terms In The Field of Information Technology Based on The Dynamic Circulation Corpus. Proceedings of IEEE.452–458.

Zhao Xiao-bing. 2007. A study on Recognition and extraction method of Contemporary Chinese Basic Vocabulary Based on Dynamic Circuit Corpus. Beijing: Beijing Language and Culture University.

Jirapa Vitayapirak, Phornsuk Ratiroch-anant. 2006. Computational Approach for Processing of Control Engineer Text: Applications for Corpus Lexicography. Proceedings of IEEE.

Han Xiu-juan. 2007. Research on Word Distribution of General Words and Relations among Characters Words and Phrases Based on Dynamic Circulating Corpus. Beijing: Beijing Language and Culture University.

Bilikezi. 2005.The application of corpus linguistics and the Uyghur corpus frequency statistical significance. Journal of Xinjiang Normal University, 26(2):226–228.

Zheng Ze-zhi, Zhang Pu, Yang Jian-guo. 2004. The Research on Lettered-word Extraction in Chinese Texts. JOURNAL OF CHINESE INFORMATION PROCESSING, 19(2):78-85.

Jirumtu, Gardi, Saiyin. 1995. Desining and Realizing the Software System of Modern Mongolian Word Frequency Statistics. JOURNAL OF CHINESE INFORMATIONPROCESSING, 11(3):24–29.

Mayra Hapar, Gulila Altenbek. 2011. Design and Implementation of Kazakh Text Categorization System. Computer Engineering, 37(5):196–198.

Wang Qiang-jun. 2003. Dynamic Circulation Corpus (DCC) Based Automatic Unlisted Term Extraction in the Field of Information Technology. Beijing: Beijing Language and Culture University.

MODIFICATIONS OF MORPHOLOGICAL ANALYSIS PROGRAMS FOR THE PROBLEMS OF MULTILINGUAL SEARCH[1]

Ayrat Gatiatullin, Madekhur Ayupov

Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan, Russia
Kazan Federal University, Kazan, Russia

This article is devoted to the description of development of integrated linguistic data models and software modules for the morphological analysis of the Turkic word forms used in the tasks of multilingual search.

Tatar, Kazakh and Turkish languages database are filled on the base of the developed models. In the process of database filling in was made a comparative analysis of the structural and functional features of different Turkic languages affixed morphemes. Due to the fact that in different Turkic languages the same morphological categories can be expressed as a synthetic and analytical method, the authors decided to develop morpho-syntactical analyzer. For the implementation of the various sub-tasks we need different modifications of the morpho-syntactical analyzer which features are described in this article.

The created model and the databases can be implemented not only in semantic search technology to enhance the functionality of search engines, but also in other Turkic languages processing systems.

Введение

Тюркские языки обладают значительными структурными отличиями от индоевропейских языков, как на морфологическом, так и на синтаксическом уровне, что обуславливает необходимость существенной доработки компьютерных моделей и программного обеспечения, разработанных для индоевропейских языков при их языковой локализации. К числу таких особенностей относятся автоматная левосторонняя морфология, агглютинация, отсутствие жесткой границы между парадигматическими классами, потенциально неограниченный объем парадигмы, нежесткое распределение лексики по грамматическим классам и частям речи. При на-

[1] Работа выполнена при финансовой поддержке РФФИ (проект №13-07-00494) Разработка комплексных моделей данных на основе ситуационного анализа текстов в задачах многоязычного поиска

личии основных структурных признаков общих для всех тюркских языков, есть признаки свойственные только определенной подгруппе тюркских языков или даже отдельному тюркскому языку.

В Институте Прикладная семиотика продолжается реализация моделей и программных модулей в рамках проекта по разработке комплексных моделей данных на основе ситуационного анализа текстов в задачах многоязычного поиска (Гатиатуллин, 2014). Одним из базовых элементов в задачах многоязычного поиска для тюркских языков является программа морфологического анализа. Несмотря на то, что известно несколько вариантов морфологических анализаторов для тюркских словоформ, одним из наиболее популярных среди которых является анализатор на основе PC-KIMMO (Koskenniemi, 1983), требуются различные модификации морфоанализаторов, удовлетворяющих определенным требованиям для использования в различных программах. К числу таких требований относится морфологический анализ без использования словаря основ или морфологический анализ с учетом аналитических единиц. Учет аналитических единиц на этапе морфологического анализа необходим по той причине, что одна и та же морфологическая категория в одних тюркских языках может выражаться с помощью аффиксов, а в других с помощью аналитических единиц. Так например, такая грамматическая категория, как инструментатив в татарском языке выражаетя с помощью послелога *белән*, в казахском с помощью аффикса –Бен, в турецком с помощью аффикса –lA, в крымскотатарском с помощью аффикса –нен. Комплексность, разрабатываемых моделей предполагает также, что источником информации для программных модулей морфологического и синтаксического анализа является единая многоязычная база данных. Эта база данных построена на основе структурно-функциональной компьютерной модели тюркской аффиксальной морфемы. Заполнение структурно-функциональной модели ведется для татарского, казахского и турецкого языков. Выбор языков обусловлен проверкой методов на примере двух более близких языков, входящих в одну кипчакскую подгруппу тюркских языков и третьего более отдаленного, входящего в огузскую подгруппу тюркских языков.

Различные программные модули, использующие информацию из базы данных многоязычной структурно-функциональной мо-

дели тюркской аффиксальной модели также служат для проверки адекватности и достоверности введенной информации.

Еще одной особенностью, разрабатываемых вариантов анализатора является то, что программная часть является универсальной для всех тюркских языков, а вся информация необходимая для анализа словоформ конкретного языка находится в базе данных.

Первая версия морфологического анализатора

Специфика первой версии морфологического анализатора в том, чтобы максимум информации заложить в Базе данных с относительно простой программной частью. Программная часть реализована в виде веб-интерфейса, который позволяет производить запросы к Базе данных. Таким образом, весь процесс морфологического анализа представляет собой поиск в базе данных элементов, удовлетворяющих заданным параметрам. База данных морфологического анализатора для каждого из тюркских языков состоит из двух словарей:

- Словарь основ
- Словарь окончаний.

В словаре основ все основы классифицированы по морфологическим и морфонологическим типам, а в словаре окончаний хранятся наборы парадигм для каждого из морфонологического типов. Окончания представляют собой цепочки алломорфов, образованные по морфонологическим правилам татарского языка. Теоретически эти цепочки, образуемые словоизменительными аффиксами, в агглютинативных тюркских языках могут иметь бесконечную длину. Однако при создании базы данных было принято ограничение по заполнению цепочек окончаний, состоящих не более чем из пяти алломорфов, что является обоснованным со статистической точки зрения.

Более подробно структура словаря описана в работе (Гатиатуллин, 2014)

Модификация морфологического анализатора

Одним из способов оптимизации работы морфоанализатора было исключение из словаря окончаний нерегулярных аффиксальных цепочек, таких которые могут присоединяться не ко всем

основам определенного типа. Например, к числу таких аффиксов в тюркских языках относятся аффиксы возвратного (–Ын) и страдательного (-Ыл) залогов. В отличие от аффиксов каузативного и взаимно-совместного залогов они присоединяются только к определенным корневым морфемам, поэтому информацию о возможности их присоединения лучше хранить в словаре основ.

Это изменение позволило сократить объем словаря окончаний, однако он все равно получился довольно большим. Этот фактор существенно влияет на время поиска в базе данных и соответственно на время анализа словоформы, в то время, когда одним из основных требований разработчиков поисковой системы Exactus является скорость работы анализатора. Основным замедлителем работы анализатора является количество обращений к базе данных и передача по сети полученной информации. Для устранения этих недостатков было решено оставить запросы к базе данных только для поиска основ, а окончания генерировать программным способом, загрузив морфотактические правила в оперативную память. Получился морфоанализатор, алгоритм работы которого напоминает алгоритм работы вышеупомянутого двухуровневого морфоанализатора PC-KIMMO. Однако, в их алгоритмах есть ряд отличий. В PC-KIMMO три типа правил: правила следования аффиксов, правила выбора алломорфа и правила фонологических изменений. В нашем варианте анализатора все три типа правил заменены одним типом: правилами следования алломорфов в словоформе.

Морфологический анализатор без словаря основ

В тюркских текстах встречается много слов, которых нет в базовых лексических словарях. Это различные имена собственные или заимствования из других языков. Строить для них отдельные словари нереально, поскольку данное слово может встретиться лишь однажды в одном конкретном тексте. С целью обеспечения возможности семантического поиска, для предложений с такими словоформами также бывает необходимо построить описание ситуации. Поскольку для выражения отношений между объектами ситуации используются аффиксальные морфемы, то для выделения этих отношений требуется определять категориальную принадлежность основы словоформы, а также набор его аффиксов.

В результате работы морфологического анализатора без использования словаря основ, количество вариантов анализа будет намного больше, чем при работе со словарем. При выдаче результатов анализа программа автоматически производит ранжирование выдаваемых вариантов анализа. Результатами с наибольшим рангом будем считать, цепочки аффиксальных морфем с наибольшим числом аффиксов.

Например:

Получая на вход словоформу *Иванныкыларга* – программа выдаст, следующие варианты анализа:

N (*Иван*) – нЫкЫ – Лар - ГА	ранг 3
N (*Иванныкы*) – Лар - ГА	ранг 2
N (*Иванныкылар*) – ГА	ранг 1
V(*Иванныкыла*) – ЫРГА	ранг 1

Механизм работы программы морфологического анализа заключается в следующем. Программа морфологического анализа проверяет возможность получения аффиксальных цепочек на основе правил следования алломорфов, а также соответствие типа, получаемой основы необходимым для используемых алломорфов морфонологическим признакам. Вся требуемая для работы программы информация находится в оперативной памяти, которая загружается при запуске программы. Таким образом, отсутствует обращение к сетевой базе данных, что способствует увеличению скорости обработки анализируемых данных.

Техническая реализация

Отдельные модели данных на основе ситуационного анализа текстов в задачах многоязычного поиска реализованы в виде клиент-серверного Web-приложения и размещены на сайте http://exact.antat.ru/. Для реализации этой системы использовалась система управления базами данных (БД) Postgresql 9.1.

В БД каждый из близкородственных языков имеет идентичную структуру (рис. 1).

Таблица ***sem*** содержит наибольшую неизменяемую форму основ для конкретного языка. Для тех основ, у которых наибольшая неизменяемая форма отличается от словарной формы, в таблице ***sem_dictform*** хранятся словарные формы основ. Таблица ***mt_cross***

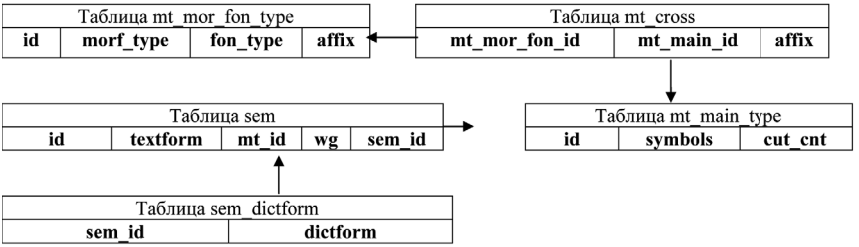


Рис. 1. Архитектура базы данных для отдельного языка

содержит всевозможные окончания, таблица ***mt_mor_fon_type*** содержит морфологические формы, таблица ***mt_main_type*** отражает морфонологические типы.

Таблицы для отдельных языков соединяются между собой через таблицу ***sem***, в которой отображаются слова русского языка.

Для внедрения прикладных морфологических моделей, словарей и баз данных татарского языка в метапоисковую систему Exactus и адаптации к ней, морфологические базы татарского языка были интегрированы в лингвистический процессор Exactus, разработанный в ИСА РАН. Также разработан модуль анализа та-

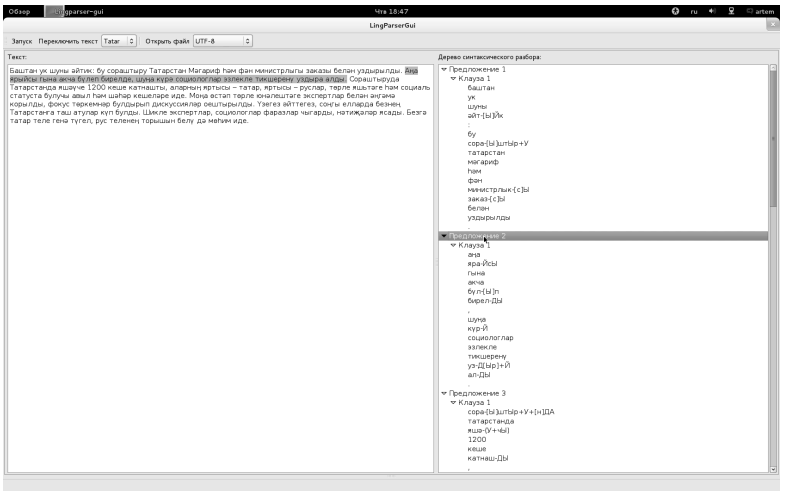


Рис. 2. Результат анализа текста на татарском языке лингвистическим процессором Exactus (просмотр анализа предложения)

тарского языка, который выполняет разбиение текста на татарском языке на слова и предложения, и для каждого слова устанавливает нормальную форму и морфологические признаки. Результаты анализа текстов на татарском языке отображаются в графическом интерфейсе (рис. 2).

Заключение

Опыт работы авторов по созданию комплексных моделей для задач многоязычного поиска на примере тюркских языков, еще раз подтверждает, что требуется большой объем работ по сравнительному анализу и унификации системы обозначений грамматических категорий и языковых единиц, используемых в тюркских языках. Результаты работы семинара UniTurk являются особенно ценным материалом для многоязычных разработок с использованием нескольких тюркских языков.

ЛИТЕРАТУРА

Гатиатуллин А.Р., Аюпов М.М. (2014). *Разработка многоязычных комплексных моделей для метапоисковой системы.* – материалы V Международной научно-практической конференции (Казань, 19–22 ноября 2014 г.). – Казань: Отечество.

Koskenniemi K. (1983). Two – level morphology: A general computational model of word-form recognition and production. Tech. rep. Publication, No. 11, Department of General Linguistics, University of Helsinki.

AN IMPLEMENTATION OF TATAR ORTHOGRAPHY USING THE NÜVE FRAMEWORK

Ercan Gökgöz ercangokgoz@gmail.com
Kalmamat Kulamshaev
kkulamshaev@fatih.edu.tr, Fatih University, Turkey
Harun R. Zafer
harunzafer@gmail.com, TUBITAK, Turkey
Samet Öztoprak, İsmet Biner, Atakan Kurt
sametoztoprak@hotmail.com ismet.biner@istanbul.edu.tr
atakan.kurt@istanbul.edu.tr, Istanbul University, Turkey

In this paper, we present a short overview of Tatar Orthography from an NLP point of view which is part of a larger ongoing Tatar-Turkish Machine Translation project we are currently working on. In this project, we model both Tatar and Turkish using the two-level morphology of Koskemnieni. The two-level morphology consists of two-level orthographic rules and finite state automatons (FSA) for morphotactics. Translation basically carried out in three main steps: (i) morphological analysis in Tatar, (ii) Tatar-Turkish translation using a dictionary, (iii) morphological generation in Turkish. We implemented the two-level rules in the Nuve Framework to create a morphological parser for Tatar. Nuve is generic two-level parser/generator framework for agglutinative languages written primarily for parsing Turkic Languages.

1. Introduction

Tatar is an agglutinative languages in Altay language group. The similarity shared by Tatar and other Turkic languages makes it easier to design and implement a translation system between Tatar and Turkish. Agglutination is very strong in Tatar which can generate variety of words and long word forms. Here is such an example:

Татарлаштырылмаганнардан(сыз)мы

The affix –сыз is put in parentheses because it is not a common place suffix nowadays. Affixes in this example are shown below:

Татар-лаш-тыр-ыл-маг-ан-нар-дан-(сыз)-мы

The dashes signify morpheme boundaries. This word could be translated into

English roughly as ‘Are you one of those whom we could not have converted to Tatar’. Vowels in a suffix must be in agreement the vowels in preceding affixes in accordance with the vowel harmony rule. Other rules requires that vowels in the roots or morphemes be removed in certain situations, as consonants are sometimes subject various phonetic rules. Tatar has many words borrowed from foreign languages such as Russian, Arabic, Persian, etc. These words usually do not obey phonetic rules.

2. Related Work

The RUSLAN project is developed to translate documents from Czech to Russian (Hadic 1987) which are both Slavonic languages. Later another translation project Cesilko was undertaken to translate from Czeck to Slovac (Hajic et al. 2000). Ceilko was extended to other Slavonic languages, Polish and Lower Serbian (Dworak et al. 2006), and the Baltic language Lithuanian (Hajic et al. 2003)

InterNOSTRUM project was another project for closely related languages. It was developed to translate between Catalan and Spanish (Canals-Marote et al. 2000). More work for Romance Language pairs was also implemented. Portuguese-Spanish machine translation system was implemented in a similar manner. (Garrido-Alenda et al. 2003)

The two-level Turkish morphology was described by Oflazer (Oflazer, 1994) using PC KIMMO (Antwort 1990) which is milestone project. In a way it paved the way for similar studies on Turkic languages, since Turkish and other Turkic languages have lexical, morphological and syntactic similarities.

The machine translation between Turkic languages is limited to only a few projects. Lexicon-based translation system was implemented between Azerbaijani and Turkish (Hamzaoglu 1993). Another early system was implemented between Crimean Tatar and Turkish (Altintas, Çicekli 2001).

A Turkmen Turkish machine translation system was implemented at Fatih University (Shylov 2008) which turned into morphological machine translation system for Turkic languages, called DİLMAÇ. Another Turkmen Turkish system was implemented at ITU (Tantug 2006,2007). We use DİLMAÇ to implement Kyrgyz Turkish translation system. Two-level morphology of Kazakh Language was described and implemented in DILMAC (Zafer, 2011). In a similar study two-level morphology of Kazan Tatar language was developed using DILMAÇ

(Gökgöz, 2011) But this study used a Latin script not the original Cyrlic script for Tatar. There were some limitations and performance concerns, so we decided to restudy and reimplement Tatar orthography and morphology in a new framework.

3. Tatar Orthography

Though in the past other scripts were used in written language, the currently Cyrillic script is used. In the table below, we give the Cyrillic script along with corresponding Latin letters.

Table 1

Cyrillic Tatar alphabet

Cyrillic	Latin	Cyrillic	Latin	Cyrillic	Latin
А а	A a	К к	K/Q k/q	Ф ф	F f
Ә ә	Ä ä	Л л	L l	Х х	X x
Б б	B b	М м	M m	Һ һ	H h
В в	W/V w/v	Н н	N n	Ц ц	Ts ts
Г г	G/G g/ğ	Ң, ң	Ñ ñ	Ч ч	Ç ç
Д д	D d	О о	O o	Ш ш	Ş ş
Е е	É é	Ө ө	Ö ö	Щ щ	Şç şç
Ё ё	Yo yo	П п	P p	Ъ, ъ	‘
Ж ж	J j	Р р	R r	Ы ы	I ı
Җ җ	C c	С с	S s	Ь, ь	‘
З з	Z z	Т т	T t	Э э	E , e
И и	İ i	У у	U/W u/w	Ю ю	Yu yu
Й й	Y y	Ү ү	Ü/W ü/w	Я я	Ya ya

The Tatar language has an alphabet consisting of 37 letters and 2 accent symbols with13 vowels: а э ә ы е и о ө у ү я ю ё and 24 consonants: й б к ф в л һ г м д н ң ч ш щ п ж җ з р с т х ц. The usual phonetic classification of vowels and consonants in Tatar is given below:

Table 2

Vowels in Tatar

	Flat				Round			
	Extensive		Narrow		Extensive		Narrow	
Thick	а	я	ы		о	ё	у	ю
Thin	э	ә	и	е	Ө		ү	

Table 3

Consonants in Tatar

Stop		Bilabial	Labio-Dental	Dental Alveolar	Palato-Alveolar	Palatal	Velar	Glottal
	voiceless	п		т	ч		к	
	voiced	б		д	җ		г	
Fricative								
	voiceless		ф	с	ш			
	voiced		в	з				
Nasal		м		н			ң	
Liquid								
	lateral			л				
	nonlat			р				
Glide						й		h

We define Tatar using two-level morphological model of Koskemnieni. In this model orthographic rules define morphophonemic process taking place 2 different levels or representations of a word: lexical level and surface level. Lexical level represent the morphemes of a word in a certain notation using meta phonems or letter classes. Phonetically similar sounds or letters which participate in certain phonetic evens are grouped together to form meta morphemes. Surface representation of a word is the word as it appears in written form normally. The key to the model is formulate the orthographic rules of a languages as changes occurs between lexical level and surface level. The meta phonems of Tatar are given below:

1. Back vowels : V_b = { а, ы, о, у, я, ю }
2. Front vowels : V_f = { э, ә, и, е, ү, ө, ё }
3. Low vowels : V_l = { а, ә }
4. High vowels: V_h = { ы , е }
5. Vowels : V = { а, э, ә, ы, е, и, о, ө, у, ү, я, ю, ё }
6. Consonants : C = { й, б, к, ф, в, л, h, г, м, д, н, ң, ч, ш, щ, п, ж, ж, з, р, с, т, х, ц }
7. Voiceless Consonants : C_s = { ф, с, т, к, ч, ш, h, п }
8. Voiced Consonants : C_t = { й, б, в, л, г, м, д, н, ң, щ, ж, ж, з, р, х, ц }
9. Nasal Consonants : C_n = { м, н, ң }

3.1. Two-Level Orthographic Rules

The rules in two-level notation are given below. The previous indicates the structural form of a word and the next indicates the phonological realization of the word. There are the examples which contain '0' symbol. It means that the symbol is removed from the surface realizations.

Two-level morphology is a general form to illustrate morphological description of word structures. Two-level orthographic rules represent phonetic events occurring during affixing of suffixes to a root or stem. The rules express the correspondence between lexical and surface forms of a word when a certain phonetic event occurs. Lexical form of a morpheme is a structural representation using meta phonems whereas the surface form is written form produced by applying the rule to the lexical form.

The colon (:) sign represents the correspondence between the lexical form on the left and the surface form on the right. The plus (+) sign represents morpheme boundary. The zero (0) symbol represents letters removed on the surface form. The rules consists of antecedent and consequent expressions. Antecedent describes a phonetic event such as *A:a* which means that a meta morpheme *A* in the lexical form changes in to letter *a* on the surface. The consequent describes the context, conditions or when the event in the antecedent should take place. The context is given in the form LC_RC. LC and RC defines the left and the right context for the event. Below we give only a few orthographic rules of Tatar because of space limitations.

1. V_l: a => ^{V}ьC* +:0*^{V}ь'@'__
2. V_l: ә => ^{V}fC* +:0 *^{V}f'@'__

The two rules above defines vowel harmony rules for back and front vowels in Tatar.

3. V_h : ы => ^{V}ьC* +:0 *^{V}ь'@'__
4. V_h : е => ^{V}fC* +:0 *^{V}f'@'__

The rules above defines vowel harmony with respect to meta morpheme V_h.

5. V_h: 0 ⇔$:0__ C +:0 [$V_l$:@ | V_h:@]

In some cases, stems have vowels which are removed when certain suffixes are affixed to them.

4. Morphophonemic processes

We go through the morphophonemic processes in Tatar in this section. We use the following meta-phonemes in defining the morphophonemic processes in Tatar which is implement as orthographic rules in the Nuve framework:

1. D : voiced (д) or lateral(л)
2. A : back (а) or front (ә)
3. H : high vowel (ы , е)
4. L : Nasal(н) or lateral(л)
5. P : voiced (п) or voiceless (б)
6. N : voiced (д) or voiceless (т) or nasal(н)
7. B : voiced (б) or voiced (г)
8. T : voiced (д) or voiceless (т)

4.1. Vowel Harmony

Tatar has vowel harmony like other Turkic languages. Vowels in a word should be in the same vowel subset according to this rule. When a suffix affixed to a morpheme, the vowel in the suffix should be in harmony with the last vowel in the morpheme.

4.1.1. Resolving low-unrounded vowels

If the last vowel of a word is a member of back vowels. The A lexical changes into a or else ә. (If the last consonant of a word is one of (м,н,ң), the L turn into ң or л)

Lexical	экран-LAp	N(screen)-PLU
Surface	экран0лар	экранлар
Lexical	адрес-LAp	N(address)-PLU
Surface	адрес0ләр	адресләр
Lexical	дошман –Lap	N(enemy)-PLU
Surface	дошман0нар	дошманнар
Lexical	көн-LAp	N(days)-PLU
Surface	көн0нәр	көннәр

4.1.2. Resolving high vowels

The lexical H is converted to ы in case preceding vowel is a back vowel, or else it is converted to е in order to comply with the vowel harmony.

Lexical	тоз-сНз	N(salt)
Surface	тоз0сыз	тозсыз
Lexical	күз-сНз	N(eye)
Surface	күз0sez	күзsez

4.2. Vowel Drops

An unstressed vowel in the middle of a word drops in case the suffix being affixed begins with e or ы vowel.

Lexical	селек-Н	N(shake)
Surface	сел$ек0е	селке
Lexical	авыз-Н	N(mouth)
Surface	авыз0ы	авзы

4.3. Vowel Changes

If the stem or root ends with one of the low vowels and the suffix starts with й letter, then the letter a at the end of the word turns into ы, and the letter ә at the end of the word turns into е.

Lexical	сырла–й	N(paint)-PLU
Surface	сырлы0й	сырлый
Lexical	үлчә-й	N(measure)-PLU
Surface	үлче0й	үлчей

4.4. Consonant Changes

If the last letter of a morpheme is C_3, and the suffix starting T is affixed to it, then T turns into т, otherwise it turns into д by default.

Lexical	як-ТАш	N(Fellow)-PLU
Surface	як0таш	якташ
Lexical	яшь-ТАш	N(coeval)-PLU
Surface	яшь0тәш	яшьтәш
Lexical	ыруг-ТАş	N(tribe)-PLU
Surface	ыруг0даш	ыругдаш
Lexical	өй-ТАш	N(Greening)-PLU
Surface	өй0дәш	өйдәш

If the first letter of the affix is either с or ч and the stem ends in з, then з changes based on the first letter of affix.

Lexical	кыз-сА	N(girl)
Surface	кыс0сА	кыссА
Lexical	кыз-чНк	N(girl)
Surface	кыч0чык	кыччык

5. Implementation in Nüve

There are several morphologic analyzers (parsers) exists for Turkish. These can be grouped by 3 types as the top-down analyzers [G. Eryiğit] [H. R. Zafer], the bottom-up analyzers [A Akın] [Hankamer] and the FST (finite-state transducer) based ones [K Oflazer] [Z Sak] [Ç. Çöltekin].

Nüve (core) is a generic top-down morphologic analyzer for Turkic Languages. It is open source and being developed with C# programming language on .Net platform. Nüve is designed to be language independent. All language specific data is defined in external files and no programming is required to adapt a new language. Nüve is especially designed for Turkic languages which are extensively agglutinative by suffixes. Nüve mainly uses 4 files as a resource for language specific data. These are

1- Root lexicon
2- Suffix lexicon
3- Orthography rules
4- Morphotactic rules

When a word is given as input to Nüve, an analysis process is performed as follows. Root and suffix lexicons include both lexical form and all possible surface forms of each root and suffix. First from left to right all the candidate roots are found. Then for each candidate, on the rest of the word surface, next suffix candidates are found. If a transition from a root candidate to suffix candidate exists, on the rest of the word surface, same process continues recursively until all the surface of the word is consumed.

The process explained above returns a list of possible solutions for the word where each solution is a sequence of morphemes. As a second process orthographic rules are processed for each solution candidate and the final surface is produces. If the final surface matches with the original word's surface, the sequence is accepted as a valid analysis of the word.

We have implemented the orthographic rules in Nuve and tested the examples given above. We have a limited root lexicon at present. However we plan to complete the lexicon as soon as possible. We have a large suffix lexicon. We plan to have extensive testing after completing the lexicons with a large sample text. We already have a Turkish analyzer and parser implement in Nuve by Harun R Zafer. In the end of this phase we will have both a parser and an analyzer for Tatar. In the next phase we will implement a Tatar morphological disambiguator and Tatar POS tagger. After that we plan to implement Tatar-Turkish machine translation system.

6. Conclusion and Future Work

An in-depth analyses of orthography and morphophonemic processes of Tatar is the starting point of successful morphological analysis of Tatar. Therefore we studied current Tatar orthography in Cyril alphabet and formulated phonetic processes using two-level orthographic rules which describes how phonetic changes are carried out when suffixes are affixed to roots or stems between lexical and surface representation of a word.

We have gone through all suffix morphemes of Tatar to formulate the two-level formalism and categorize them, according to the usage in morphology. Then we went on to study each morpheme to which morphophonemic processes they go through, to reach a more complete description of Tatar orthography and morphology.

Finally, we implemented the two-level rules in the Nuve Framework to create a morphological parser and generator for Tatar. Nuve is generic two-level parser/generator framework for agglutinative languages written primarily for parsing Turkic Languages.

REFERENCES

Koskenniemi K. (1983). Two – level morphology: A general computational model of word-form recognition and production. Tech. rep. Publication, No. 11, Department of General Linguistics, University of Helsinki.

Karttunen L (1983). PC-KIMMO: A General Morphological Processor. In Texas Linguistics Forum 22, pp.165–186.

Oflazer, K. (1994). Two-level description of Turkish morphology, Literary and Linguistic Computing, Literary and Linguistic Computing Volume9, Issue2 pp. 137–148.

Shylov M. (2008). DİLMAÇ: Turkish and Turkmen Morphological Analyzer and Machine Translation Program. Master's thesis, Fatih University, İstanbul Turkey.

Gökgöz E., et al (2011). Two-Level Qazan Tatar Morphology. 1st International Conference on Foreign Language Teaching and Applied Linguistics.

Zafer H.R, et al (2011). Two-Level Description of Kazakh Morphology. 1st International Conference on Foreign Language Teaching and Applied Linguistics.

Altintaş K & Çicekli İ. (2001) A Morphological Analyser for Crimean Tarar. In: Procedings of the 10th Turkish Sym on AI and NN TAINN pp 180–189, North Cyprus.

Canals-Marote R., et al. (2000) interNOSTRUM: a Spanish-Catalan Machine Translation System. Machine Translation Review, 11, 21–25

Tantuğ A.C., Adalı E. & Oflazer K. (2006) Computer Analysis of the Turkmen Language Morphology. In: FinTAL, Lecture Notes in Computer Science, pp. 186–193. Springer

Dvořák B., Homola P. & Kuboň V. (2006) Exploiting similarity in the MT into a minority language. In: LREC-2006: Fifth International Conference on Language Resources and Evaluation. 5th SALTMIL Workshop on Minority Languages, Italy

Garrido-Alenda A., et al. (2003) Shallow Parsing for Portuguese-Spanish Machine Translation In: TASHA 2003: Workshop on Tagging and Shallow Processing of Portuguese, Lisbon, Portugal

Hajič J. (1987) RUSLAN – An MT System Between Closely Related Languages. In: Third Conference of the European Chapter of the Association for Computational Linguistics (EACL'87), Copenhagen, Denmark

Hajič J., Homola P. & Kuboň V. (2003) A simple multilingual machine translation system. In: MT Summit IX, New Orleans, USA

Hajič J., Hric J. & Kuboň V. (2000) Machine translation of very close languages. In: Proceedings of the sixth conference on Applied natural language processing pp. 7–12. Morgan Kaufmann Publishers Inc., Proceedings of the sixth conference on Applied natural language processing

M.Corbi-Bellot A., et al. (2005) An open-source shallow-transfer machine translation engine for the Romance languages of Spain. In: 10th EAMT conference "Practical applications of machine translation", Budapest, Hungary

G. Eryiğit and E. Adalı, "An affix stripping morphological analyzer for Turkish," in Artificial Intelligence and Applications: IASTED International Conference Proceedings, as part of the 22 nd IASTED International Multi-Conference on Applied Informatics, 2004.

H. R. Zafer, "Nuve: A Natural Language Processing Library for Turkish

in C#" [Online]. Available: https://github.com/hrzafer/nuve. [Accessed: 05-Jul-2015].

Akın, M. D., & Akın, A. A. (2009). Zemberek - Açık Kaynak Kodlu Türkçe DDi kütüphanesi [Online]. http://code.google.com/p/zemberek/. [Accessed: 05-Jul-2015].

Hankamer, J. (1986). Finite state morphology and left to right. Fifth West Coast Conference on Formal Linguistics, (s. 29-34). Stanford.

K. Oflazer, "Two-level description of Turkish morphology," Lit. Linguist. Comput., vol. 9, no. 2, p. 137, 1994.

H. Sak, T. Güngör, and M. Saraçlar, "A Stochastic Finite-State Morphological Parser for Turkish," aclweb.org, no. August, pp. 273–276, 2009.

Ç. Çöltekin, "A Freely Available Morphological Analyzer for Turkish," Proc. 7th Int. Conf., pp. 820–827, 2010.

COMPUTER ASSISTED LANGUAGE LEARNING: A CRITICAL EVALUATION, INTRODUCTION, HISTORY AND ACHIEVEMENTS

Saringul Ziyadova

Azerbaijan National Academy of Sciences, Istiqlaliyyet str. 30, Baku, AZ1001, Azerbaijan

Computer Assisted Language Learning (CALL) is a widely researched and discussed field in language learning. The amount of literature is impressive: there are thousands of published articles. This paper provides a general introduction on CALL – its history, researches and achievements in the field of ICALL so far. Introduction part includes the reasons for computers and CALL's entrance to the world of instruction and the definitions for the CALL phenomenon various researchers provided. Next paragraph talks about the history of CALL. That is the stages the CALL went through with regard to the speedy and unexpected developments. Those stages are referred to as models in the paper and are discussed in properly and in details. History part is followed by further researches part where the latest researches and their results are introduced. Those researches aim to find answers to critical questions such as "Is CALL better than traditional instructions?" or "Is CALL that much useful as an electronic instructor?" The research work about ICALL (Intelligent Computer Assisted Language Learning) is also included in this paper. ICALL section is divided into two parts. First part talks about definitions for ICALL by several researchers and the reasons for ICALL being introduced separately from CALL, while the second part introduces the best language teaching software so far and discusses them in details. This is basically a critical review and analysis and review of various works done in this field so far. As the phenomenon of CALL is quite recent and new in Azerbaijan, there is not much work and research done in this regard. Therefore, references used in the paper include foreign authors.

Introduction and the review of interpretations of CALL

In former times the speed of life was noticeably law, as the technology had not taken over the world's trends and demands, yet. The speed of every field, as well as education corresponded to that trend. In other words, there was no need to operate faster, as no competent rival existed to cope with. Nevertheless, technological boom brought a demand of an insane speed to operate in social, economic, political and individual lives. Variety of disciplines launched researches to find

ways to operate quickly, productively and effectively, while keeping quality as a top priority. Taking this strive under consideration, production of microcomputers with the rapid development of information technologies brought convenience to every field of social life, economy as well as education. The occurrence of the internet and the ease of the access to the World Wide Web even globalized computers. Park and Son (2009) state, "The internet, particularly, has become a useful tool for communication, a venue for experiencing different cultures and a mediator in diverse political, social and economical situations" (2009:1). Thus, the instructors and linguists, trying to be perfect in language teaching, decided to proof check and implement the aforementioned wonderful tools and be creative in their methods of language teaching, as well. The technological means breaking cultural and linguistic boundaries and getting people around the world closer ideally fitted the instruction purposes, as interaction and communication are the primary features of language learning. Although some other means of technology (e.g., TV, gramophone, radio, tape-recorders etc.) were used in language teaching/learning, computers appeared to be the most successful and convenient facilities in everyday school life. Computer Assisted Language Learning (CALL) refers to as implementing software, internet opportunities, audio and video files, global integration via the internet to language learning. Levy (as cited in Gruba, 2004) defines CALL as "the search for and the study of applications of the computer in language teaching and learning". According to Rahimi and Yadollahi (as cited in Talebi&Teimuri, 2013) CALL is usually understood as "an approach to language teaching and learning in which the computer is used as an aid to the presentation, reinforcement and assessment of the material going to be learned" (2013:52). Hashemi and Aziznezhad (as cited in Talebi&Teimuri, 2013) stated that one of the biggest pros of CALL is that it underpinnes the process of autonomous learners generation. They believed that before applying computer in the process of language instruction, the instructors should consider several crucial factors; evaluation of the learners' computer skills to make them aware of the basic usages of the computer, the language capacity and web navigation skills of the learners, and some technical issues which should be taken into account such as access to network environment, use of modern equipment and software, being informed of elementary internet technology, and major and possible problems by teachers.

As Houizhong believes (as cited in Ifeoma, 2010) "CALL is when the computer is being used as an instructional tool to improve learning content" (2010:66). The opinion of Kearsley (as cited in Ifeoma, 2010) is "It includes the use of simulations, drills, tutorials, word processing, authored programmes, games, database search/inquiry methods and programmed instruction" (2010:66). Januszewski and Molenda (as cited in Ifeoma, 2010) interprets CALL as "… a technique for using technology in the field of language learning" (2010:67). Davies (as cited in Mahdi, 2013) defines CALL as "an approach to language teaching and learning in which computer technology is used as an aid to the presentation, reinforcement, and assessment of material to be learned, usually including a sustainable interactive element" (2013:192). Davies (as cited in Mahdi, 2013) classifies two major types of computer application in language instructions and learning. They are (1) the ones that involve the use of generic software tools such as word processors, presentation software, e-mail packages, and web browsers; and (2) the ones designed particularly to promote language learning. The use of computers in the field of language learning is studied under the area of CALL. Pacheco (as cited in Chiu, 2003) argues that based on the behavioristic (see *History* chapter for more about behavioristic theory) theories of Bloomfield and Skinner the first development of CALL was an electronic extension of programmed learning or programmed instruction. His arguments on the theories of Bloomfield and Skinner (as cited in Chiu, 2003) were as follows:

> According to these theories, all learning could be broken down into small "frames" and the learner could be drilled and evaluated in frame until mastery. The teacher then brought the student to the next frame. In the computerized version, the progress of the student could be monitored and guided through "branching". Proficient students could automatically be sent ahead, while slower students could be routed to remedial lessons (2003:5).

Seljan, Banek, Špiranek and Lasić-Lazić (n.d.) described the latest model of CALL as an integrative one "where computer is not only used as a media for delivering instructions as in a behavioristic phase or as a tool in a communicative phase, but integrates multimedia packages, CD-ROMs and internet supporting skill-based activities, interactive learning and self-access as an approach in teaching and learning" (n.d.:1).

Overall, Computer Assisted Language Learning is a discipline where language teaching and learning is attempted to be integrated with technology while keeping instructors assistance actual. This field reflects the synthesis of traditional language teaching methods and computer and internet related tools to broaden the borders of classroom into global environment and real-life language atmosphere.

History, models of CALL, further researches and ICALL

1. History

The assistance of computer in education first took place in 1950s for other purposes than language teaching (Tafazoli & Golshan, 2014). Richard Atkinson and Patrick Suppes applied the well-known early CALL project at Standford University, US (Chapelle, 2001; cited in Tafazoli & Golshan, 2014). Atkinson (as cited in Tafazoli & Golshan, 2014) marked that the project, in collaboration with IBM, was based on his mathematical learning theory, but not on language learning theory. Despite the fact that various gadgets and technologies were used in language teaching and learning, the first grandiose project was built in 1960s and was referred to as Programmed Logic for Automated Teaching Operations (PLATO). In other words, the period of 60s was the starting point in revolutionary changes of approaches to language learning and a challenge to traditional language instructors. Gruba (2004) noted that PLATO system was developed at the University of Illinois to enable teachers to write a Russian-English translation course. The computer program was able to provide drills and marking for student work as well as an authoring component for instructors (Gruba, 2004). In 1980 PLATO system was used to instruct English, French, German, Spanish and Italian languages (Hendricks, Bennion & Larson, 1983; cited in Tafazoli & Golshan, 2014). Levy (as cited in Gruba, 2004) underlines that "the Time-Shared, Interactive, Computer-Controlled Information Television (TICCIT) project initiated at Brigham Young University in 1971 as one of the first examples of multimedia-based instruction". In that system learners were able to control audio, video, and text integrated to computers. TICCIT was designed on instructional

grounds and allowed instructors to add content. Nevertheless, it did not allow deciding on how to teach with the programmed materials (Levy, 1997; cited in Gruba, 2004). Paramskas (as cited in Tafazoli & Golshan, 2014) stated that the Computer-Assisted Learning Exercises for French (CLEF) was initiated by three Canadian Universities to instruct basic French grammar. Paramskas and Thomas (as cited in Heimpel, 1998) noted that CLEF was one of the collaborative works of language instructors and computer programmers. Levy (as cited in Gruba, 2004) also talked about Athena Language Learning Project initiated at the Massachusetts Institute of Technology (MIT). He noted that a significant part of the project was the full integration of language teachers in the development process. In other words, project managers promoted the application of software design and instructional theory in teaching and learning.

The area of discipline integrating language learning and computer technology is referred to as Computer Assisted Language Learning (CALL) and the term is unanimously accepted, although earlier researchers labeled their searches in the related field with acronyms such as CAI (computer-aided instruction), CAL (computer-assisted learning), CELL (computer-enhanced language learning), and TELL (technology-enhanced language learning) (Gruba, 2004). The term CALL was agreed upon in TESOL convention held in Canada. Tafazoli and Golshan (2014) stated about the convention as follows:

> The 1983 TESOL convention in Canada was the milestone in CALL from two aspects: 1. The CALL was the expression agreed upon. 2. A suggestion was made to establish a professional organization titled "CALICO" (Computer-Assisted Language Instruction Consortium). By that time CALL flourished in education and market settings: a course on CALL at Lancaster University, EuroCALL professional organization, production of introductory materials, and publication of a large number of books (2014:33).

2. Models of CALL

According to the choices of application and methods of use the approaches to this field are divided into three models: *Behaviouristic* CALL (*Structural* CALL), *Communicative* CALL and *Integrative* CALL (Warschauer, 1996; cited in Staszika, 2007).

2.1. Structural CALL

The structural model took a start in 1950s and "drill and practice" was the leading motto of that model (Warschauer, 1996; cited in Staszika, 2007). The wider use of structural CALL started in the 1960s and 1970s, as audio-lingual teaching method was commonly implemented in language instruction (Seljan, Berger & Dovedan, 2004). Atkinson and Wilson (as cited in Tafazoli & Golshan, 2014) classified three major factors that affected the use of CALL: "(a) the use of programmed instruction based on behaviorism, (b) the enhanced sophistication of data processing, and (c) the use of time-sharing system for CALL purposes" (2004:33). Computer in language learning was perceived as a tireless and patient assistant to practice exercises reiteratedly. Grammar-translation and audio-lingual methods were the leading teaching methodologies at that time and computers were adapted to serve the corresponding methods. In other words, *computer as a tutor* (Taylor, 1980; cited in Staszika, 2007) was the basic model in structural CALL. Taylor (as cited in Tafazoli & Golshan, 2014) emphasized that computer operated as a tutor, that is the materials and the exercises it presented were repetitive language drills, vocabulary, grammar, and translation tests. An approach of this kind was also beneficial for students with law speed and capacity of language learning. Considering the style of delivering materials, PLATO was the best system to represent structural CALL. Bangs and Cantos (2004) noted that the expensive PLATO project expanded because of the failure of educational system in the United States. The reason for it was the hope for the possibility to improve the educational state. The terms such as "instructional model" (Philips, 1987; cited in Bangs & Cantos, 2004) and "wrong-try-again" model (Underwood, 1984; cited in Bangs & Cantos, 2004) were also mentioned, as the attendance of teachers was not necessary. Crook (as cited in Gruba, 2004) defines two reasons why tutorial drills were so attractive to language instructors: 1. Teachers, viewing new technology as a challenge, were pretty much comfortable with automatic and user-friendly programmes. 2. Educators considered repetition of practices and continual exposure to certain tests truly useful.

Dina and Cironei (as cited in Tafazoli & Golshan, 2014) suggested many beneficial sides of repetitive language drills and practices:

1. Providing whenever necessary access to the same learning material is essential to acquiring a language;

2. Allowing students to access the same material over and over again and offering immediate and non-judgmental feedback every time is ideal for mastering a language;

3. Presenting such language materials on an individualized basis, without time keeping and deadlines, offering students the choice to study in their own rhythm is beneficial for owing a language (2014:33).

Nevertheless, CALL became no more demandable for two grounds (Bangs & Cantos, 2004):

1. The lack of imagination and creativity in designing new and challenging exercises;

2. The high cost and maintenance of the computers (2004:224).

2.2. Communicative CALL

Communicative model of CALL started in 1970s and 1980s and changed individual exposition to unnatural language patterns into integration into real-life language patterns. This model also turned CALL into more interactive discipline which allowed interaction between learners. The real-life language patterns helped students to be aware of how real communication with native speakers would look like and how to respond and interact effectively. Singhal (as cited in Tafazoli & Golshan, 2014) mentioned that technology provides language learners to be acquainted with the culture of native speakers of the language. That is, they can later be aware of the way the cultural background may impact one's view of the world (Singhal, 1997; cited in Tafazoli & Golshan, 2014).

John Underwood was the initiator of Communicative CALL model in late 1970s and early 1980s. Warschauer (as cited in Staszika, 2007) noted that John Underwood offered a series of "Premises for 'Communicative' CALL" in 1984 (Underwood, 1984; cited in Staszika, 2007). The premises include as follows:

- Focusing more on using forms rather than on the forms themselves;
- Teaching grammar implicitly rather than explicitly;
- Allowing and encouraging students to generate original utterances rather than just manipulating prefabricated language;
- Not judging or evaluating everything the students uttered nor rewarding them with the congratulatory messages, lights, or bells;
- Avoiding telling students they are wrong and being flexible to a variety of student responses;

- Using the target language exclusively and creating an environment in which using the target language feels natural, both on and off the screen;
- Never trying to do anything that a book can just do as well (2007:12).

According to Lee (as cited in Chiu, 2003) the software used in communicative model were *text reconstruction programmes* and *simulation*. Taylor and Perez (as cited in Tafazoli & Golshan, 2014) used the term *stimulus* for communicative model.

Communicative and behaviorist approaches had similarities to some extent. That is, both included exercises and tests. Higgins and Johns (as cited in Tafazoli & Golshan, 2014) stated that the activities and instruction based on text reconstruction and different types of cloze exercises were considered to be communicative. Chapelle's description (as cited in Tafazoli & Golshan, 2014) of communicative exercises was as follows:

> …variations included: words deleted on a fixed-ratio basis, words deleted on the basis of some criteria, or all words deleted, texts that the teacher entered into the program, or texts other learners constructed; with help options and scoring, or with simple yes/no judgments concerning the correctness of learner's entries; with the end result begin the completed text, or the end result to comprehension questions about the text (2014:34).

2.3. Integrative CALL

Eased internet access and improved multimedia opportunities of computers globalized the interaction demand of communicative model. Language learners gained the access to almost native environment of their target language just by pressing a few buttons of PC. Thus, internet made it possible to bring the whole world inside one small room within a few seconds, instead of travelling to native speakers' country which was expensive and time-consuming. Furthermore, language educators broadened the frame of language instruction to integrating four skills – reading, writing, listening, and speaking, into instruction.

Authentic project works and *meaningful interactions* in CALL (Gruba, 2004) are mostly implemented features of integrative model. Students here are encouraged to make researches and produce their own works (Gruba, 2004).

Integrative CALL is not fully researched and classified as other two fields. Unlike structural and communicative CALL, integrative CALL

is not limited within the frame of classroom activities. The introduction of World Wide Web, publishing techniques and tools, and opportunities to communicate all around the globe provided the chance to develop language skills independently and beyond the frame of instructor's syllabuses.

Wyatt's study (as cited in Chiu, 2003) proposed that CALL was beneficial and appealing as the instruction with that method expected to be "1) interactive, 2) active, 3) informative, 4) self-paced, 5) student-centered, 6) neutral, and 7) patient" (2003:38).

According to Warschauer (as cited in Staszika, 2007) CALL is still on the stage of development and striving to assist a student in semantic issues, in other words to reach the level of Intelligent CALL, where he notes that CALL program

should ideally be able to understand a user's *spoken* input and evaluate it not just for correctness but also for *appropriateness*. It should be able to diagnose a student's problems with pronunciation, syntax, or usage and then intelligently decide among a range of options (e.g. repeating, paraphrasing, slowing down, correcting, or directing students to background explanations) (2007:14).

3. Further researches and experiments

Computer Assisted Language Learning requires further improvements and researches, as gaps under discussion and disputes still exist. CALL programs can only operate within the frame of their database, while language learning requires an imaginative and creative approach. Unlike grammatical errors, it is beyond the capacity of the computers to recognize logical errors. However, many conducted researches show that CALL is more effective than traditional language instruction in terms of time, real-life language patterns exposure, accuracy, fluency, and comprehension.

3.1. Researches

Kılıçkaya (2005) conducted a research among two groups, namely control and experimental group, 17 participants per each. Each group was trained on the basis of TOEFL exam program within 8 weeks. Control group was instructed in a traditional method, while experimental group was trained in a language laboratory with the computer assisted instruction method. The results on the overall grade and competence

were almost similar, no remarkable difference was identified. However, the difference between the scores in reading and listening sections were noteworthy. Experimental group showed better results in them.

Berns, Rodriguez, and Gomez (2013) conducted a pilot study to analyze the opportunities of 3-D online games provided in language acquisition and communication of the students at A1 level. The 16 students from the University of Cadiz were asked to play two games named *Hidden room-game* and *Shopping-game. Shopping-game* was considered to be more dynamic and faster, while *Hidden room-game* was perceived as more useful in terms of language improvement according to individual feedback from the questionnaires held among the students. As a result, the students improved their writing skills and vocabulary, "solved language problems by paraphrasing, question rising, as well as clarification and confirmation requests" (2013:25).

Barani (2013) conducted a research project to investigate the effect of CALL on vocabulary achievement of Iranian students. Randomly selected 72 students were divided into experimental and control groups. Both groups were taught unknown words in 10 sessions. Experimental group member were taught with the assistance of language software, while control group members were taught in a traditional method. The end result revealed that use of CALL is more helpful in vocabulary improvement, as experimental group members outperformed.

Chenu, Gayraud, Martinie, and Wu (2007) investigated the effectiveness of CALL on learning of French relative clauses. Participants of the research were randomly selected among the intermediate non-native users of French and were divided into control and experimental groups. Experimental group members also were expected to fill in the questionnaire to give feedback on CALL. The end results showed that no significant difference occurred between the results of both groups and traditional method was preferable for participants. Nevertheless, low-level participants were more progressive in the experimental condition than those in the traditional condition. To that end, the use of CALL in grammar teaching is more effective for less proficient students and traditional method does not suit every students need.

4. ICALL

ICALL is an abbreviation for Intelligent Computer Assisted Language Learning and started to be treated as a separate field of research

ten years ago. The term intelligent addresses the criticisms toward the traditional limited and unimaginative drill exercises of structural and communicative CALL models. Despite the fact that integrative CALL model provided the learners with the opportunity to go beyond the borders of classroom environment and reach the native target language environment simply by moving fingertips, the highly motivated learners prefer well-organized curriculum when learning a language. Integrative CALL is chaotic and unorganized in nature and still requires deep research and classification, although this model perfectly fits for remarkably talented learners striving to break the limits of traditional instruction and outperform their peers in their own pace. Intelligent CALL became a separate field after Artificial Intelligence (AI) was developed enough to be added to language learning systems (Shaalan, 2005). Shaalan (2005) also noted that "the beginning of the new research field was characterized by Intelligent Tutoring Systems (ITS), which embedded some NLP features to extend the functionality of traditional language learning systems" (2005:2). Comparing ICALL to business software, Warschauer and Healey (1998) highlighted that "the best new business software offers users help at every step; good CALL software should do no less" (1998:9). They considered the simple answers of software, namely 'right' or 'wrong, try again', less helpful and emphasized that an 'intelligent' software would be the one responding not only over correctness or incorrectness of the answer but also indicate the reason of the result and propose resources for further information.

John Underwood (as cited in Warschauer & Healey, 1998) in his work named 'On the edge: Intelligent CALL in 1990s' hoped for the software which would allow students to ask the software for help if any problem with regard to NLP occurred. He hoped for progress in the fields of artificial intelligence, hypermedia and simulations so that language learning models with the assistance of CALL discipline was improved (Warschauer & Healey, 1998).

According to Warschauer and Healey (1998) the major goal in the field of ICALL is the software's intelligent response to the language learners output. They use the term of *natural language processing* for the process. However, Amaral and Meurers (n.d.) mentioned that ICALL projects have very little influence on foreign language programs, as issues important for the design of ICALL projects such as activity design, language assessment and measurement, teaching techniques, syllabus

design, second language analysis, cognitive models of second language acquisition, and language policy and planning are beyond the area of expertise of computational linguists and computer scientists.

Generally speaking ICALL is expected to reach the level of artificial intelligence and linguists and computer scientists are striving to reach it. Moreover, significant development is observed on the way to achieve the desired level. Current highly developed language learning and instruction projects of ICALL are the best indicators for that development.

4.1. ICALL projects

The major list of successful ICALL projects is as follows: *Robo Sensei* (Nagata, 2002; as cited in Amaral & Meurers, n.d.)*, E-Tutor (Heift, 2010), Lärka* (Volodina, Pilan, Borin & Tiedemann, n.d.), *Spanish for Business Professionals* (Hagen, 1999; as cited in Amaral & Meurers, n.d.), *TAGARELA* (Amaral & Meurers, n.d.), *and Oahpas (*UiT Norgga árktalaš universtehta/UiT The Arctic University of Norway*,* 2014).

Robo Sensei system was constructed by Noriko Nagata to teach Japanese language to foreigners and it consists of 24 lessons. This system perfectly fits instructors' preferences as a classroom e-coursebook. Each activity in the system is accompanied by visual aids with pictures of Japan or Japanese drawings (Amaral&Meurers, n.d.). According to Nagata (2005) *Robo Sensei* lessons contain 5 types of exercises per each, i.e. (1) vocabulary and grammatical pattern, (2) noun and verb phrases, (3) sentence composition, (4) reading comprehension, and (5) dictation. Although the system is designed to teach Japanese, the language of instruction, i.e. descriptions of exercises, feedback on exercises, terms of use, user interface is in English. The reason for such an extensive use of English is the non-Roman alphabet of Japanese. Moreover, beginner learners may be discouraged to learn the Japanese if they come across with too much complexities (Amaral&Meurers, n.d.).

E-Tutor system was developed by Trude Heift and first implemented in 1999 in Simon Fraser University to teach German language (Heift, 2003 as cited in Amaral&Meurers, n.d.). This system overcame several improvements over a decade and consists of 15 chapters. Each of those chapters starts with the introduction text which shortly describes the contents of the chapter. The system offers nine different activity types to develop listening and reading comprehension, culture, and writing (Heift, 2010). According to Heift (2010), "*E-Tutor* provides more tradi-

tional learning environments in which learning activities are performed individually" (2010:445).

Spanish for Business Professionals (SBP) is specifically designed to teach business Spanish, unlike previously mentioned ICALL systems. SBP is particularly appealing for the perfect selection of audio materials, visual aids, additional links to grammar explanations and the words directly linked to an electronic bilingual dictionary. Exercise types of it are vocabulary exercises, translation exercises, *charadas,* reading comprehension, and dictation. *Charadas* is a particular exercise where students are asked to identify the letters in a word and put the words in a correct order (Amaral&Meurers, n.d.).

TAGARELA (*Teaching Aid for Grammatical Awareness, Recognition and Enhancement)* is designed to teach Portuguese by complementing already existing pedagogical materials of introductory kind of Portuguese. The system provides exercises adapted to practice reading, listening and writing skills. TAGARELA offers six types of exercises, namely reading comprehension, fill-in-the-blanks, listening comprehension, vocabulary, rephrasing, and picture description. All the activities are provided with the immediate and individualized feedback (Amaral&Meurers, n.d.).

Oahpa! And Oahpa!-nuõrti: Oahpa! is software for interactive training of vocabulary, grammar and communicative skills. It was first designed for North Sámi by the staff of the Center for Sámi Language Technology at the University of Tromsø. Later on, new versions for Sámi and non- Sámi languages have been initiated and are under creation. *Oahpa!-nuõrti* is specifically designed for Skolt- Sámi languages spoken in Finland, Norway and Russia (UiT Norgga árktalaš universtehta/UiT The Arctic University of Norway, 2014)

Conclusion

Already 50 years have passed since 1960s, when first attempts to integrate computer to language learning were made. Further developments in information technologies, production of micro-computers and well-structured and more user-friendly software challenged linguists and programmers to frequently switch attitudes towards CALL and develop new and flexible strategies to achieve the top target. The top target indeed was to make computers useful for development of four skills (i.e. listening, speaking, reading, and writing) in language learning. Intel-

ligent CALL systems have been constructed allowing extremely busy people learn a language, thanks to the development of artificial intelligence. Although ICALL is only useful in the improvement of listening, reading and writing skills, the social network is the best solution to practice speaking with foreigners. Internet and social network are the key terms to practice the language, have access to real native language materials and be exposed to native language without a need to live in the native speakers' country. Education becomes more accessible and free of charge every passing day, as an access to internet was made cheaper. The websites like www.coursera.org and other analogues ones are the best examples on the fact that one does not need to face bureaucratic obstacles or deal with tons of papers and forms to get the desired education. As technology develops at an insane speed one can never claim that our today's perceptions about CALL would stay concrete and unchanged in the future. The developers of PLATO system also anticipated that PLATO would become in consistent use all over the world, while later profound changes and growth in information technology completely revolutionized the world wide perceptions about CALL. Information Technologies is rapidly growing field and any spontaneous revolutionary invention must be expected, although computers already took its deserved place in language learning and instruction processes.

REFERENCES

Amaral, L. A. & Meurers, D. (n.d.). *On using intelligent Computer-assisted language learning in real-life foreign language teaching.* Retrieved December 13, 2013 from Universität Tübingen website: http://www.sfs.uni-tuebingen.de/~dm/papers/amaral-meurers-11.pdf.

Bailey, S. M. (2008). Content assessment in intelligent Computer-aided language learning: Meaning error diagnosis for English as a second language. (Doctoral dissertation, Ohio State University, 2008). Retrieved December 13, 2014 from the Ohio State University website: https://etd.ohiolink.edu/rws_etd/document/get/osu1204556485/inline.

Bangs, P. & Cantos P. (2004). What can Computer assisted language learning contribute to foreign language pedagogy? *International Journal of English Studies,* 4(1), Retrieved December 12, 2014 from **dialnet**.unirioja.es/descarga/articulo/919602.pdf.

Barani, G. (2013). The impact of Computer assisted language learning (CALL) on vocabulary achievement of Iranian University students EFL

learners. *International Journal of Basic Sciences & Applied Research,* 2(5), (pp. 531–537) Retrieved December 12, 2014 from http://isicenter.org/fulltext/paper-118.pdf.

Berns, A., Rodriguez F. & Gomez R. (2013, July). Collaborative learning in 3 D virtual environments. Paper presented at the WorldCALL 2013- Global perspectives on Computer-assisted language learning conference, Glasgow, UK. Retrieved December 12, 2014 from University of Ulster website: http://www.arts.ulster.ac.uk/worldcall2013/userfiles/file/shortpapers.pdf.

Chenu, F., Gayraud, F., Martinie, B. & Wu, T. (2007). Is CALL efficient for grammar learning? *JALT CALL Journal, 3(3), (pp. 85–83).* Retrieved December 12, 2014 from http://journal.jaltcall.org/articles/3_3_Chenu.pdf.

Chiu, M. W. (2003). Computer-assisted language learning: Attitudes of Taiwanese college students (Doctoral dissertation, University of West Florida, 2003). Retrieved December 12, 2014 from http://ir.csu.edu.tw/bitstream/987654321/619/1/00212939.pdf.

Gruba, P. (2004). Computer assisted language learning (CALL). In Davies A. & Elder C. (Eds.), *The handbook of applied linguistics* (pp. 642–667). Retrieved December 12, 2014 from https://www.academia.edu/1121087/The_Handbook_of_Applied_Linguistics.

Heift T. (2010). Developing an intelligent language tutor. *CALICO Journal,* 27(3), (pp. 443–459). Retrieved December 13, 2014 from https://calico.org/html/article_811.pdf.

Heimpel, R. (1998). *The multimedia network-based language learning centre: A historical approach.* Retrieved December 12, 2014 from http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/61/94.

Hubbard, P. (n.d.). *General introduction.* Retrieved December 12, 2014 from the University of Stanford website: http://web.stanford.edu/~efs/callcc/callcc-intro.pdf.

Ifeoma, O. E. (2010). Using Computer-assisted language learning to improve student's English language achievement in universal basic education. *International Journal of Research and Technology,* 1(1), (pp. 66–71). Retrieved December 13, 2014 from http://soeagra.com/ijert/vol1/ijert9.pdf.

Kılıçkaya, F. (2005). *The effect of Computer-assisted language learning on learners' achievement on the TOEFL exam.* Published master's thesis. Middle East Technical University, School of Social Sciences, MA. Retrieved December 13, 2014 from http://etd.lib.metu.edu.tr/upload/12606252/index.pdf.

Mahdi, H. S. (2013). Issues of Computer assisted language learning normalization in EFL contexts. *International Journal of Linguistics,* 5(1), Retrieved December 12, 2014 from http://www.macrothink.org/journal/index.php/ijl/article/viewFile/3305/pdf.

Nagata, N. (2005). *Robo-Sensei personal Japanese tutor: Features & benefits.* Retrieved December 13, 2014 from the University of San Francisco website: http://usf.usfca.edu/japanese/RSdemo/preRSfiles/features.htm.

Park, C. N. & Son, J. B. (2009). Implementing Computer-assisted language learning in the EFL classroom: Teacher's perceptions and perspectives. *International Journal of Pedagogies and Learning,* 5(2), Retrieved December 13, 2014 from https://www.academia.edu/1365822/Implementing_computer-assisted_language_learning_in_the_EFL_classroom_teachers_perceptions_and_perspectives.

Seljan, S., Berger, N. & Dovedan, Z. (2004). *Computer-assisted language learning (CALL).* Retrieved December 12, 2014 from the University of Zagreb for Information Sciences website: http://dzs.ffzg.unizg.hr/text/call.pdf.

Seljan, S., Banek, M., Špiranek, S. & Lasić-Lazić, J. (n.d.). *CALL (Computer-assisted language learning) and distance learning.* Retrieved December 12, 2014 from University of Zagreb for Information Sciences website: http://dzs.ffzg.unizg.hr/text/call_dl.pdf.

Staszika, S. (2007). CALL software – Experimental study (Doctoral dissertation, Instytut Humanistyczny, 2007). Retrieved December 13, 2014 from http://www.cs.put.poznan.pl/ksiek/_publications/sie07.pdf.

Tafazoli, D. & Golshan, N. (2014). Review of Computer-assisted language learning: History, merits & barriers. *International Journal of Language and Linguistics,* 2(4), (pp. 32–38). Retrieved December 13, 2014 from http://article.sciencepublishinggroup.com/pdf/10.11648.j.ijll.s.2014020501.15.pdf.

Talebi, F. & Teimuri, N. (2013). The effect of Computer-assisted language learning on improving EFL learners' pronunciation ability. *World Journal of English Language,* 3(2), (pp. 52–56). Retrieved December 13, 2014 from http://webcache.googleusercontent.com/search?q=cache:EktSHSubIGAJ:www.sciedu.ca/journal/index.php/wjel/article/download/2953/1744+&cd=1&hl=az&ct=clnk&gl=az.

UiT Norgga árktalaš universtehta/UiT The Arctic University of Norway. (2014). *About the project Skolt- Sámi Oahpa!-nuõrti.* Retrieved December 13, 2014 from http://oahpa.no/sms/useoahpa/background.eng.html#Oahpa%21.

Volodina, E., Pilan, I., Borin, L. & Tiederman, T. L. (2014, May). A flexible language learning platform based on language resources and web services. Paper presented at LREC conference, Reykyavik, Iceland. Retrieved December 13, 2014 from http://www.lrec-conf.org/proceedings/lrec2014/pdf/892_Paper.pdf.

Warschauer, M. & Healey, D. (1998). *Computers and language learning: an overview.* Retrieved December 13, 2014 from http://hstrik.ruhosting.nl/wordpress/wp-content/uploads/2013/03/Warschauer-Healey-1998.pdf.

http://smallseotools.com/plagiarism-checker/.

IMPROVING MORPHOLOGICAL ANALYZER: THE COMPLEX NETWORKS-BASED APPROACH

Denis Kirjanov, Boris Orekhov

Higher School of Economics, National Research University
Moscow, Russia

This study is aimed at improvement of the morphological analyser Bashmorph (Орехов 2014) by the means of complex networks based on affixal chains taken from the analyses provided by the parser. We compiled a complex network where the nodes represent affixes and the edges between them represent their coocurrences. The edges have weight, which is equal to the number of cooccurrences.

Parameters of such a network reflect various language phenomena; but they can also reveal errors of the analyser. For instance, loops in the network indicate morph recursion which is impossible in Bashkir and therefore they point at mistakes of analysis. Edges between affixes which contain vowels of different rows is most often provoked by wrong analysis of Russian words like именно as a form of Bashkir word им 'medicine'. Finally, as it's revealed by the means of the graph, variability in affix ordering is strongly restricted so it seems reasonable to set all the possible affix orders in analyser in advance.

В 2014 г. в публикации (Орехов 2014) был представлен автоматический анализатор башкирской морфологии bashmorph. После представления программы работа над ним была продолжена. На материале размеченного корпуса были построены графы, которые помогли выявить рад закономерностей, позволивших улучшить алгоритм разбора.

Нашими данными послужил корпус, составленный из текстов статей газеты «Йәшлек» ('*молодость*') за 2007–2014 гг. Суммарный объём корпуса – 5,8 млн словоупотреблений. Тексты газеты были размечены при помощи морфологического анализатора bashmorph (далее – парсера; разработчик – Б. В. Орехов, см. подробнее (Орехов 2014)). Каждому слову приписывалась морфологическая аннотация, но при этом грамматическая неоднозначность не снималась, в анализаторе отсутствует соответствующая функциональность. Структура разметки на выходе программы устроена следующим образом: сначала приводится словоформа, затем –

лемма, потом следует аффиксный состав, далее – перевод леммы на русский язык, после чего читатель находит глоссы для каждой морфемы исходной словоформы; в случае, если теоретически возможно несколько разборов, то они разделены знаком вертикальной черты. Символы оформления вывода программы в основном следуют формату, предусмотренному в морфологическом анализаторе для русского языка “Mystem” [Segalovich 2003].

Пример (1) – это пример данных на выходе работы программы:

(1) *йыл.{йыл+Ø=»год»=S=NOM,SG|йы+л=»собирать»=V=PASS}*

Нашим материалом являлись все возможные разборы (так, из примера (1) мы использовали оба разбора). В ходе предварительной обработки корпуса мы извлекли из каждого возможного разбора только цепочки аффиксов. Далее мы построили сеть, вершинами в которой являлись аффиксы; ребро между аффиксами появлялось в том случае, если был найден хотя бы один разбор, в котором они следуют друг за другом и между ними нет никакого другого аффикса. У каждого ребра есть вес – количество примеров совместной встречаемости двух аффиксов.

Рассмотрим пример (2):

(2) *бала-лар-ым-ды*
child-PL-POSS.1SG-ACC
‘моих детей’

При обработке такого разбора программой, генерирующей сеть, в графе строилось два ребра: *лар-ым* и *ым-ды*. Если таких рёбер до этого в графе не было, то вес каждого ребра был бы равен единице; если такие рёбра существовали до обработки этого разбора, то к весу каждого из них добавлялась единица.

Для дальнейшего изложения необходимо учесть ряд нюансов:

1) Парсер предлагает нулевое окончание (обозначается как ∅) в именительном падеже единственного числа и повелительном наклонении единственного числа;

2) В парсере существует трехуровневая иерархия разборов: разборы третьего уровня – это такие, при которых основа после разделения слова на аффикс и гипотетическую основу не обнаружена в словаре; разборы второго уровня – аналогичная ситуация,

при которой гипотетических аффиксов выделено несколько; разборы первого уровня – “уверенные” разборы, при которых основа в словаре парсера нашлась. Если парсер предлагал разбор более высокого уровня, то остальные разборы отбрасывались.

3) В начале был построен ненаправленный граф (т.е. рёбра не имели направлений).

В процессе работы мы построили несколько сетей. В первую очередь необходимость построения нескольких сетей обусловлена тем, что некоторые характеристики актуальнее смотреть на такой сети, в которой каждый алломорф каждого аффикса представлен как самостоятельная вершина, а некоторые другие – на такой, в которой вершинами являются только морфемы, а алломорфы одной морфемы представлены как единое целое.

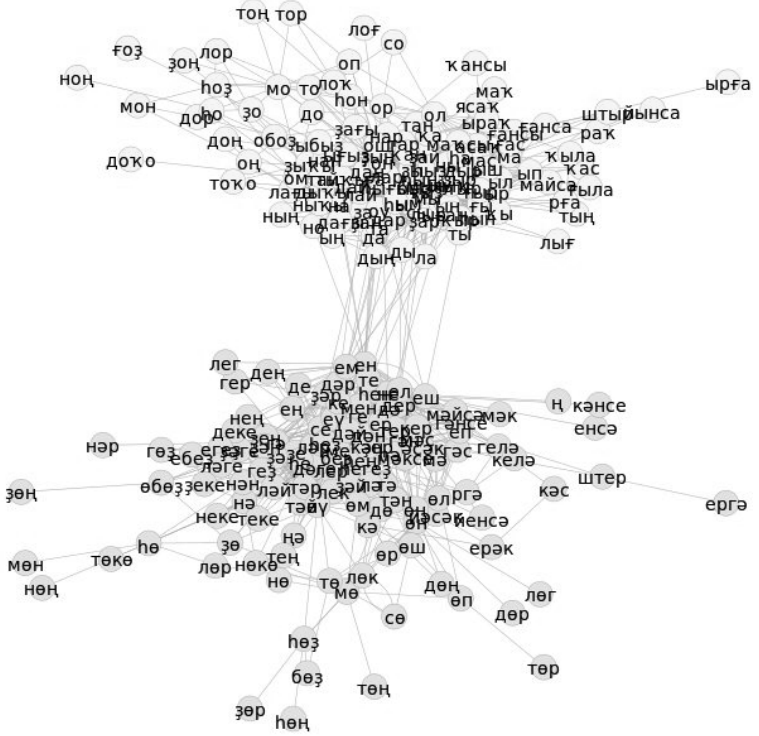


Рис. 1. Сеть совместной встречаемости аффиксов в башкирском языке.

В результате обработки данных мы получили следующие результаты. В сети 279 вершин.. При анализе этой сети обнаружилось, что в ней имеются петли (“Ребра вида (a, a) или {a} называются петлями” (Мельников 2010: 16)): ребро от некоторого аффикса вело к нему самому. Это означает, что парсер разобрал некоторые словоформы таким образом, что некоторые аффиксы повторялись в словоформе, при этом непосредственно следуя друг за другом, что невозможно в башкирском языке. Рассмотрим пример (3):

(3)*Аҡмулла*{аҡмул?+л+а=»?»=V=PASS,PRES|аҡму?+л+л+а=»?» =V=PASS,PASS,PRES}

Как мы видим, башкирское имя *Аҡмулла* разбирается как глагольная форма, второй из вариантов разбора которой включает в себя два показателя пассива *-л-*. Все примеры на 19 циклов, обнаруженных нами в построенной сети, были неверными разборами[1] (чаще всего в верном разборе этих словоформ должна была фигурировать основа длиннее той, в которой предлагался повтор аффиксов; эта основа, соответственно, оканчивалась на то же буквосочетание, которым далее выражался настоящий аффикс). В 11 из 19 случаев фигурировали однобуквенные аффиксы: можно заключить, что разборы, содержащие такие аффиксы, должны были иметь меньший вес относительно разборов, которые их не содержали.

В графе, вершинами которого являлись морфемы, также были обнаружены петли. В первую очередь речь идёт о глаголах, содержащих два показателя каузатива: такие глаголы достаточно хорошо изучены, см., например, (Kulikov 1999). Кроме того, мы обнаружили один пример с двойным реципрокальным показателем. Хотя носители башкирского языка подтверждают, что этот пример действительно верен (впрочем, исследование причин использования и семантики двойного показателя реципрока требуют отдельного исследования, наши информанты говорили о подчеркнутой длительности и взаимности процесса), говорить о двойном маркере реципрока приходится с большой осторожностью: (Ахмеров 1958: 630) помимо глагола *һорау* ‘спрашивать’ выделяет как отдельную лексическую единицу и глагол *һорашыу* ‘расспрашивать’ (возможно, исторически реципрок от *һорау*):

[1] Мы благодарны А.А. Галлямову за экспертную оценку этих разборов.

(4) *Үткән-һүткән тукта-п, күре-ше-п, хәл-әхүәл*

Туда-сюда остановиться-Ger смотреть-RECP-GER самочувствие

һора-шы-ш-а-быҙ.

спрашивать-RECP-RECP-PRS-1PL

‘Проходя мимо друг друга, мы останавливаемся, смотрим и спрашиваем, как дела’

Таким образом, можно заключить, что обнаруженные в графе, вершинами котороого являлись морфы, а не морфемы, петли, со стопроцентной вероятностью указывают на ошибки парсера.

На рис. 1 можно увидеть, что в графе выделяются две половины: морфы, содержащие гласные переднего и непереднего ряда соответственно. Тем не менее, законы сингармонизма нарушаются: эти две половины оказываются связаны. Мы проанализировали рёбра, соединяющие ту половину графа, которая содержит аффиксы с гласными переднего ряда, с той половиной, которая содержит аффиксы с гласными непереднего ряда. Таких рёбер – 25. Приведём их список в таблице 1.

Таблица 1

Связь переднерядных и непереднерядных аффиксов

Аффикс 1	**Аффикс 2**	**Частотность**	**Словоформы**	**Верный ли разбор**
те	*мы*	3	*системы, тотемы*	нет
дәр	*ҙа*	1	*көндәрҙа*	да
ләр	*ҙан*	1	*матдәләрҙан*	да
ен	*да*	14	*аренда*	нет
ем	*да*	1	*Эдемда (этемда)*	нет
ен	*на*	11	*антенна*	нет
ен	*дар*	1	*мартендарҙың*	да
ем	*дар*	2	*Гилемдар (илемдар)*	нет
ел	*сы*	1	*табелсы*	нет
та	*мен*	1	*Эштамен*	нет

еш	*ты*	1	*бүлешты*	да
ын	*дә*	2	*вакытындә, һарайындә*	да
ен	*дағы*	1	*ансамблендағы*	да
дер	*ҙар*	1	*ордерҙар*	нет
ел	*ды*	1	*Орелды*	нет
ем	*ды*	1	*Артемды*	нет
ған	*дәр*	1	*кыумағандәр*	да
ен	*дың*	2	*Андерсендың*	нет
ем	*дың*	1	*Артемдың*	нет
ен	*но*	26	*именно*	нет
еш	*оу*	5	*телешоу*	нет
ла	*мен*	2	*парламентер, парламен*	нет
он	*не*	1	*Сталонне*	нет
ош	*ен*	1	*Ярошен*	нет
кыр	*ҙәр*	1	*тапкырҙәр* (на самом деле корень *тапкыр*, а не *тап*)	нет

Прежде всего напомним, что средняя частотность пары морфов в построенной нами сети равна 1756, в то время как в таблице 4 нет ни одной пары, чья частотность превысила бы 26; невысокая частотность этих пар, помимо прочего, позволила нам отразить в столбце “словоформы” таблицы 4 все примеры из нашего корпуса, найденные на каждую из представленных в таблице пар морфов. Таким образом, можно сказать, что все эти пары достаточно периферийны и окказиональны. Помимо этого, 18 из 25 разборов неверны: как правило, ошибки парсера связаны с тем, что словоформа, заимствованная из русского языка, разбирается как башкирская словоформа, основа которой находится в словаре, см. (5):

(5)*именно{им+ен+но=»лекарство»=S=POSS.3SG,ACC|имен+но=»целый|здоровый»=ADJ=ACC|им+ен+но=»сосать»=V=PASS,PST.DEF}*

В этом примере в двух из трех возможных разборов русское наречие *именно* разобрано как форма башкирской лексемы *им*, в составе которой фигурируют аффиксы *ен* и *но*, имеющие гласные разных рядов.

Всё это говорит о том, что к парсеру должен быть подключён словарь русскоязычных слов, используемых в башкирских текстах в ситуации code-switching. Однако правила, которое запрещало бы сочетания морфов, содержащих гласные разных рядов, по-видимому, не требуется: 8 пар, представленных в таблице 1, являются следствием правильных разборов программы.

Мы анализировали граф, построенный на порядке аффиксов в словоформах, и обнаружили, что вариаивность в языке фиксируется только в трёх точках: PL-POSS1SG, PL-POSS2SG, CLIT. INDEF-CLIT.INTERROG. Таким образом, можно заключить, что для парсера «выгоднее» задать возможный порядок аффиксов, заложив в нём три точки вариативности, чем не задавать никаких правил порядка аффиксов (см. схожую идею в (Ермолаева 2015)

Таким образом, ряд обнаруженных нами свойств графа (в частности, рёбра между двумя половинами графа, содержащими морфы с гласными разных рядов) помог улучшить морфологический анализатор для башкирского языка. Помимо этого, запрет на свободный порядок аффиксов улучшает точность разборов.

Список глосс: ACC – аккузатив, CLIT.INDEF – клитика со значением неопределенности, CLIT.INTERROG – клитика со значением вопросительности, POSS.1SG – притяжательный показатель первого лица единственного числа, POSS.2SG – притяжательный показатель второго лица единственного числа, PASS – страдательный залог, PST.DEF – определённое прошедшее время, PL – множественное число, RECP – реципрок, S – существительное, V – глагол.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Ахмеров 1958 – К. З. Ахмеров (ред.). Башкирско-русский словарь (ок. 22000 слов). Государственное издательство иностранных и национальных словарей.

2. Ермолаева 2015 – М. Б. Ермолаева Прототип универсальной системы автоматического анализа морфологии для тюркских языков / Дипломная работа; Московский государственный университет. – Москва, 2015

3. Мельников 2010 – Ю. Б. Мельников. Элементы теории графов. Екатеринбург.

4. Орехов 2014 – Б. В. Орехов. Проблемы морфологической разметки башкирских текстов // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2014. – Казань: Изд-во «Фэн» Академии наук РТ. – Сс. 135–140.

5. Kulikov 1999 – L. Kulikov. Remarks on double causatives in Tuvan and other Turkic languages // In: Journal de la Société Finno-Ougrienne 88. Helsinki. Pp. 49–58.

6. Segalovich 2003 – Ilja V. Segalovich. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search. // Proceedings of MLMTA-2003. Nevada.

A STUDY ON CROSS PROCESSING BETWEEN THE SAME FAMILY AND SIMILAR LANGUAGES[1]

Muheyat Niyazbek[1,2], Kuenssaule Talp[3], Dawa Idomucao[1,2]

[1]School of Information Science and Engineering Xinjiang University, 830046, Urumqi
[2]Xinjiang Laboratory of Multi-language Information Technology, 830046, Urumqi
[3]College of Chinese Medicine of Xinjiang Medical University, 830011, Urumqi

The extension of natural language processing systems, such as MT (machine translation), to a new language especially the minority languages being used in China requires large amounts of paralleled data. Generally, it is prohibitive to collect huge paralleled data. This paper first investigates the similarity level between the same family and closer languages (such as Altai family languages) and then examines a transformation between their words and texts. Cosine similarity measure and dynamic programming (DP) algorithm are used to calculate the similarity between the source and target languages using a multilingual parallel data set. Test data set includes 7,854 paralleled sentences of Chinese, Uyghur, Kazakh and Mongolian various writing systems. Experimental results demonstrate that the similarity level of the languages from the same language branch is higher than that between different language branches while lower than that between the Altai family languages. Furthermore, we find that a text transformation of word to word units is feasible for the same language branches when replacing the function affixes of word using a function affixes rule bank to create common models. Additionally, it functions well without MT and order changes of parallel sentences.

1. Introduction

Language can be considered as a unique representation of the national culture and plays an important role in the research of the development and expanding traditional culture.

[1] Foundation item : Xinjiang Uygur Autonomous Region Science and Technology Support Project (201291116) and (61163030) ; Multi-Lingual Information Technology Laboratory of Xinjiang open projects funded (XJDX0905-2013-3)

Biography : Muheyat Niyazbek (1967-), man, associate Professor; Corresponding author : I.Dawa(1956-) ,man,Doctor, Professor; Kuenssaule Talp(1969-) ,woman ,associate Professor; E-mail: muheyatn@126.com

Researches work on the national language processing for minority languages are gradually move from character display to intelligentized text processing. However, the resources, such as bilingual corpus, are very important to the multilingual text and speech processing. It is prohibitive to accumulate and create huge number of resources for developing country-wide languages like minority languages spoken in China.

Some languages, like Altai family languages, are very similar in their writing system and spoken style. Figure 1 shows the sentence alignment examples writing by different graphic characters of Mongolian. They are used today in Inner Mongolia, Xinjiang, and Mongolia respectively. We call them Mongolian language branch [1] (MLB). Another set of sentences shown in figure 2 are Uyghur, Kazakh and Kyrgyz languages, all written by Arabic graphic characters belonging to Turkish. These languages are called TLB. We can see that each sentence consists of a sequence of entries, where two entries are separated by a space. The SOV grammar, the order of subject, predicate, object and verb, edit rule of a word and word order, are similar to each other. MGL and TLB are different in the writing style, but they are very close in SOV grammar and syntax.

We can also find the relationship between each entry indicated by bias parts in Fig. 1 or Fig. 2. In a case of MLB, a transformation of word by word or stem by stem units is derived from ToDo to NM. However, it is rather difficult for the case of TM to ToDo and NM to MGL as shown in Fig. 3. Notably from Fig. 2, a text transformation between TLB of word or stem units may be possible.

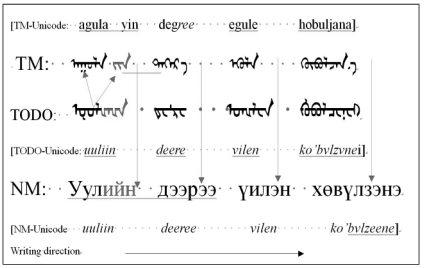


Fig.1. Example of the alignment sentences by MLB

Fig.2. Example of the alignment sentences by TLB

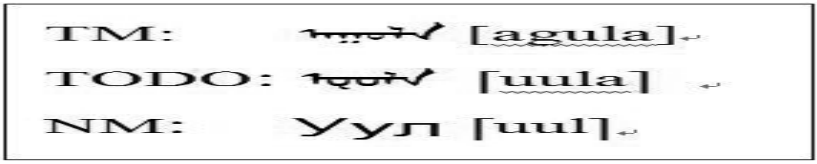


Fig. 3. Word alignment by Unicode of MGL

When performing MT from Chinese to Uyghur or to Kyrgyz, the method illustrated in Fig.4 may be an efficient way. The overall procedure is that an MT from Chinese to Uyghur is first performed, and followed by TT (text transformation) from Uyghur to Kyrgyz or Kazakh, and not Chinese to Uyghur and again do Chinese to Kyrgyz. Many common stems and the same order are also observed in the case of Uyghur to Kyrgyz or Kazakh, as indicated in red parts in Fig.4. This method can also be applied for the transformation between MLB. In this work, we will first investigate the similarity level between the texts of MLB or TLB. Then we propose an approach to transform texts between similar languages in different groups of MLB or TLB based on the similarity of bilingual text entries and DP algorithm.

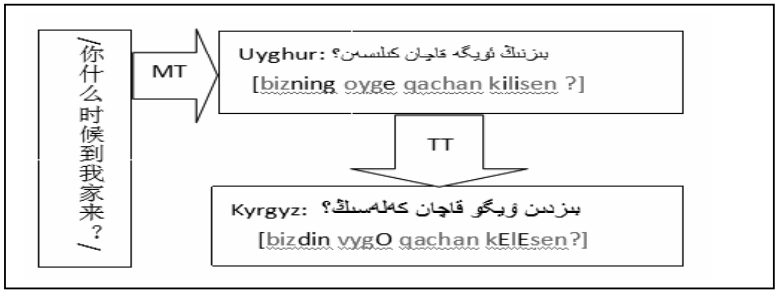


Fig. 4. An approach of MT, and then TT in a case of from Chinese to TLB

2. Previous Reaches

T.Schultz et. al. developed a way for acoustic modeling cross languages using global phoneme mapping to 15 languages acoustic data[2]. They examine what performance can be expected in this scenario. Their experiments are only for speech recognition on the resource sparse lan-

guages. The paper [3] introduces various approaches adopted in Chinese –English Cross language information retrieval via MT technique. Furthermore, there are many researches focused on only computing sentence similarity on the same language text for improving MT quality [4].

At present, researches on cross language transformation in the same family language, particularly minority languages, are not usual. In our framework, we proposed an approach based on a data-driven and linguistic rule to transform the Mongolian texts writing by multiform characters [5].

3. Investigation approach

3.1. Bilingual text similarity

In order to derive similarity of two sequences from bilingual alignment text, a cosine similarity method is used in this research.Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude [6]. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1]. The cosine of two vectors can be derived by using the following formula: Given two vectors of attributes for example, *A* and *B* (see function (1)), the cosine similarity, *cos(θ)*, is represented using a dot product and magnitude as function (2). In Here *n* and *m* is indicate the size of entries of two texts respectively.

$$\begin{cases} A = a_1, a_2, \ldots a_n \\ B = b_1, b_2, \ldots b_m \end{cases} \tag{1}$$

$$simlarity \quad = Cos\,(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{m} (B_i)^2}} \tag{2}$$

For text matching, the attribute vectors *A* and *B* are usually the term frequency vectors of the documents. The cosine similarity can be seen as a method of normalizing document length during comparison.

3.2. DP matching

In this section, a way to create the target language word model from the result of similarity word's list is proposed based on the DP algorithm. This means that to develop a rule to find a target language entry when having source language entry and similar strings resulted from alignment sentences.

DP, also known as dynamic time warping (DTW), was introduced for non-linear time alignment of two continuing patterns [7]. DP can effectively minimize errors that occur during the time alignment of the two patterns. Compared with conventional methods of matching two strings such as edit distance and longest common subsequence [8], DP is more effective because an entry can correspond to more than one entry during the matching. The DP algorithm is addressed as follows:

Consider two entries W_1 and W_2 with arbitrary length, say, l and k respectively in equation (3).

$$\begin{cases} W_1 = a_1, a_2, \ldots a_k \\ W_2 = b_1, b_2, \ldots b_l \end{cases} \tag{3}$$

Taking distance $d_t(i, j)$ between the entries, we initialize them as follows:

$$d_t = \begin{cases} 1 \quad if \quad a_i = b_j \quad (4) \\ 0 \quad Otherwise \end{cases}, \quad g_t[0][0] = \begin{cases} 0 \quad if \quad S_1[0] = S_2[0] \\ 1 \quad Otherwise \end{cases} \tag{5}$$

$$For \quad j = 1 \quad to \quad |W_2|, \quad g_t[0][j] = g_t[0][j-1] + d_t[0][j] \tag{6}$$

$$For \quad i = 1 \quad to \quad |W_1|, \quad g_t[i][0] = g_t[i-1][0] + d_t[i][0]$$

Then, the matching between entries W_1 and W_2 is regarded as a temporal alignment in a two-dimensional plane (see figure 5). Suppose that the sequence of matched pairs $c_l(i_l, j_l)$ of W_1 and W_2 forms a time warping function F expressed as,

$$F = c_1, c_2, \cdots, c_l, \cdots, c_L \tag{7}$$

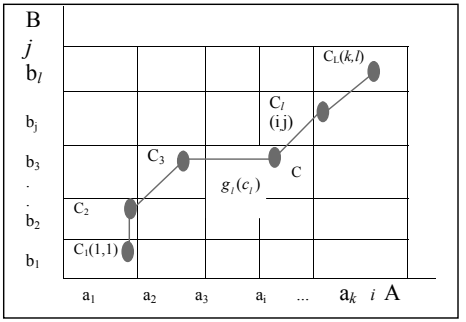


Fig. 5. An example of DP matching path

Let $g_l(c_l)$ denote the minimized overall distance representing the explicitly accumulated distance from $c_1(1,1)$ *to* $c_l(i,j)$. Then, $g_l(c_l) = g_l(i,j)$ can be expressed by equation (8) when the initializations are given by equations (5) and (6) (here, l is the size of the entries).

$$g_l(i,j) = \min\begin{cases} g_l(i,j\text{-}1) + d_l(i,j) \\ g_l(i-1,j-1) + 2\times d(i,j) \\ g_l(i-1,j) + d_l(i,j) \end{cases} \tag{8}$$

Now, if, for example, there are q candidate words to be selected and the minimized overall distance is given by $D^q_{\min}(W_1,W_2) = (1/k+l)g_q(k-1,l-1)$, then the word will finally be selected by equation (9).

$$D_x = \min\{D^q_{min}(W_1, W_2)\} \tag{9}$$

Note that, the implementation of equation (8) runs in O(k,l) time.

4. Experiments and results

4.1. Data

Table 1 test data

language	#of sentence	#of entry	#of stem	#of alphabet
Chinese	7854	101,235		
Uyghur	7654	45,872	8626	31

Kazakh	7854	38,050	7508	33
Kyrgyz	7854	34,545	8658	32
TM	7854	31015	12833	31
TODO	7854	32420	9531	31
NM	7854	28043	8416	

Focusing on the multilingual medical service system, a multilingual parallel database was built in our framework, supported by NS-FCP[SFXP] This database was collected from Chinese sentences in the medical science and medicine information domain, and translated to Uyghur, Kazakh and Mongolian manually. There are 7,854 sentence pairs used to evaluate the similarity between two language pairs. The details of test data are listed in table 1.

4.2. Similarity investigation

Experiments to investigate the similarities between language pairs were further tested in three levels, including Unicode sentence and word levels by applying the function (2) addressed in section 3 above.

First, converting the sentences to vector sequence using a space between entries was performed, and then the similarity by language pairs was computed. Figure 6 shows the results of the similarity in sentence level for MLB and TLB respectively.

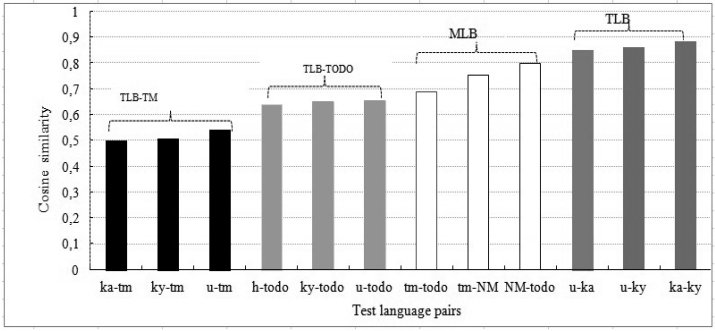


Fig. 6. Results of the sentence similarity by language pairs

Figure 6 shows some interesting phenomena. Firstly, in the case of pairs of TLB, the similarity is very high (close to 0.87), for those in the same family and called different languages. Secondly, in the case of MLB, the similarity between TM and ToDo pair is less than TLB and to 0.72 even MLB is in the same family and same nationality languages. Finally, the similarity between the different groups and used in the same areas, e.g TLB-ToDo in Fig.6, is less than MLB comparing to TLB, and higher than TLB-TM (here TLB and TM are used in Xinjiang and Inner Mongolia respectively).

Next, vocabularies are extracted from text sentences of each language, and then the similarities are computed by word set pairs. Figure 7 show the comparison of the results. We can easily see that the similarity level is higher when a word–level comparison in the same language branch while it is so far in the case of the different language branch.

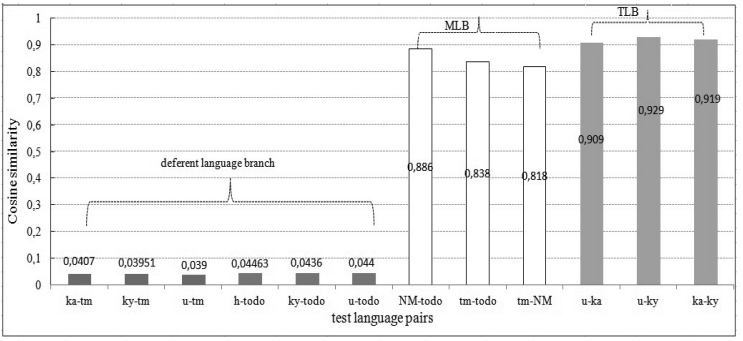


Fig.7. Results of the words similarity by language pairs

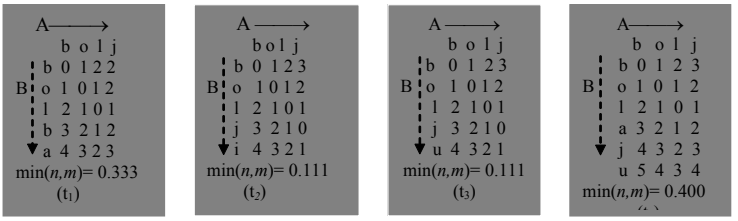


Fig. 8. DP matching bya entry “bolj” to its similar words

4.3. Extracting the target language words

In this section, we investigate a way to extract the target language words when a set of similar words from the source language for an entry of the target language applying the similarity measure. DP algorithm addressed section 3.2 above is applied to choose the target language word.

Figure 8 shows some examples for test entry "*bolj*" from MLB. There were four candidates such as tests (t_1), (t_2),(t_3) and (t_4) in figure 8. The best choose was (t_2) and (t_3) because it gives the minimal overall distance, min (*n,m*)= 0.111.The results indicate that, an entry "*bolj*", which is a verb in the source language NM word set, will be created to a word "*bolji*" of the target language ToDo when add a postfix "*i*" to source word suffix in a case of MBL. There are many similar examples of TLB, *e.g* shown Fig.4, when a function sub-word of target language Kyrgyz "*go*" will be replaced by a function sub-word "*ge*" of the source language Uyghur. Thus, it is easy to have a text transformation by word to word unit between the similar languages, such as Altai family languages without MT technique.

5. Conclusion

In this paper, we discussed a conversion method among texts for similar languages. Firstly, we investigated the similarity levels of languages in the same group and between deferent groups using cosine similarity measure. Then we investigated a way to find the target language words from the source language based on the DP algorithm

We confirmed that the transformation is more feasible by word to word units when learning the connection rule of a stem and an affix (function words) between the source and target languages by word level. Thus, this avoids the uphill work of MT for the resource-deficient languages such as minority languages being used in the developing countries. Additionally, the costs can be reduced.

In the future, we will further investigate the connection rule of a stem and an affix (function words) between the source and target languages. Meanwhile, we will work on building the statistic models for them.

REFERENCES

[1] Shi-Kuo Chang, Jong-Hyeok Lee and Kam-Fai Wong, "computer processing of orintallanguages", 2006, Vol.19(2&3), World Scientific.

[2] T.Schultz and A.Waibel, "Fast Bootstrapping of LVCSR System with Multilingual phoneme Sets, Proc. Eurospeech, 1, pp. 371–374.

[3] Lin jun Zhang, et.al, "Cross-Language information retrival", Journal of Computer cience,2004,Vol.31(7), pp. 16–19.

[4] Shen gwei Tian, et.al, "A method fro Uyghur Sentence Similarity Computation", Journal of Computer Engineering and Application, China, 2009,Vol49(26), pp. 144–146.

[5] Idomucogiin xxx and Satoshi Nakamura, "A Study on Cross Transformation of Mongolian Language", Journal Natural Language Processing, 2008, Vol.15(5), pp 3–21.

[6] Jun. Ye, "Cosine Similarity measures for intuitionistic fuzzy sets and their Applications". Mathmatical and computer Modeling, 2011 Vol.53, pp. 91–97.

[7] Landau G. M, et al, "Two algorithms for LCS consecutive suffix alignment", In Proc.15th Ann. Simp. On combinatorial Pattern Matching, LNCS 3109, 2004, pp-173–193.

[8] Francois Nicolas, Eric Rivals, "Longest common subsequence problem for unoriented and cyclic strings" J.Theoretical Computer Science 370, 2007, pp. 1–18.

EXPERIENCE OF CREATION OF LINGUISTIC SOFTWARE IN UZBEKISTAN

Anvar Nuriev

фирма ‘SPELLS’
Tashkent, Uzbekistan

This article describes the experience of creating linguistic software in Uzbekistan, as an example of software products created by the company «SPELLS» for the period from 1999 to 2005. Briefly describes the following software products:

1. Russian-Uzbek and Uzbek-Russian translation dictionary.
2. Module editing of keyboard layouts for the Uzbek language.
3. Transliterator from Uzbek (Cyrillic) to Uzbek (Latin) and vice versa.
4. Spellchecker program for Uzbek language in Microsoft Office applications.
5. Automatic translator of texts from Uzbek to Russian and vice versa.

Also in the article listed the sequence of creation and the creation of conditions for these software products, describes the difficulties developers of this software encountered while writing it. A brief review of similar softwares of companies which worked in this area in the past and are working at present.

1. Введение

Развитие технологий автоматического перевода в России и в мире, конце 90ых годов прошлого века, позволило существенно увеличить скорость работы с документами. Что в свою очередь сказалось на увеличении производительности труда. К сожалению, в это время разработок в данной области, в Узбекистане не велось. Программные продукты, создаваемые в это время не могли считаться полезными в силу слишком малой словарной базы и слабой проработки интерфейса пользователя. Данная ситуация и натолкнула на мысль о создании электронной версии словарей узбекско-русского и русско-узбекского направлений. Проанализировав опыт создания словарей русско-английского и других направлений, после изучения существующего рынка в этой области, было принято решение о начале работ по созданию словаря.

2. Электронный переводной словарь русско-узбекского и узбекско-русского направлений.

В 1999 году компания «SPELLS» выпустила первую версию электронного словаря узбекско-русского направления. За основу был взят бумажный аналог однотомного словаря 1988 года выпуска. Словарь был отсканирован и полученный файлы были обработаны с помощью автоматических средств распознавания текста. Далее текст редактировался в специальном редакторе, где слово и словарная статья(перевод) связывались между собой. Общее количество словарных статей составило около 38 тысяч. Применение тестового формата в качестве БД позволило отказаться от установки специализированных СУБД, что сделало установку программы простой и понятной.

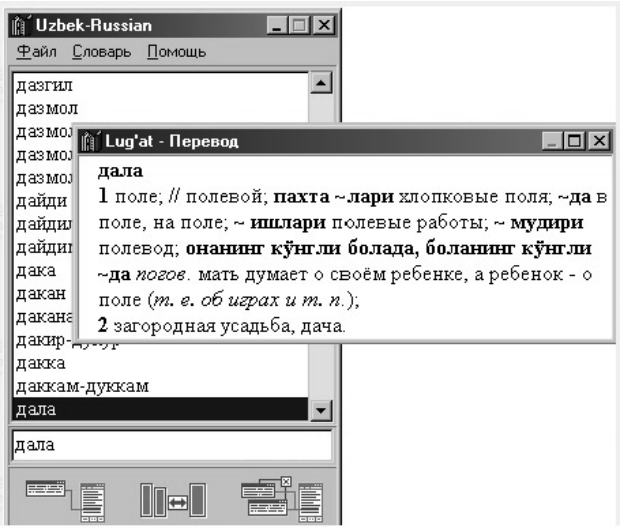


Рис 1. Узб-Рус словарь

Далее в течение года по такой же схеме был создан русско-узбекский словарь. В данном случае за основу был взят двухтомник 1983–1984 года выпуска. В процессе работы был доработан инструментарий для создания словаря в сторону удобства работы корректоров. Были попытки доработать инструментарий для кол-

лективной работы с одной БД, но в процессе работы, после расчета трудозатрат на создание подобной схемы было принято решение отказаться от такого подхода из-за больших трудозатрат.

Общее количество словарных статей в данном словаре получилось около 75 тысяч. Данные словари, как и последующие были объединены в одном интерфейсе, с возможностью выбора нужного направления перевода. Со временем были добавлены несколько словарей меньшего объема по следующим направлениям-делопроизводство, медицина, горное дело и т.д.

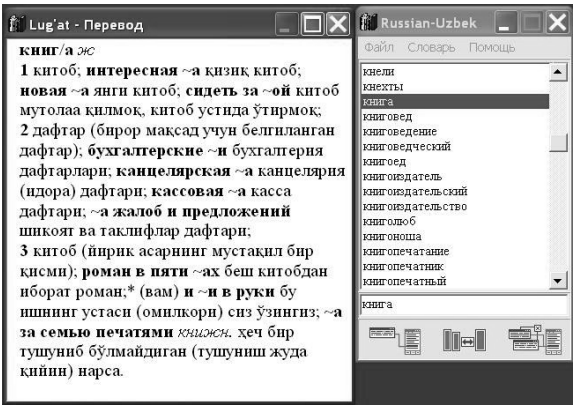


Рис 2. Рус-Узб словарь

3. Модуль редактирования раскладки клавиатуры для узбекского языка.

В силу большой распространенности операционных систем серии Windows 95-98 для полноценной работы со словарями в плане ввода 4ех символов узбекской кириллицы (“Қ”, “Ў”,” Ғ”, “Ҳ”) пришлось создать редактор клавиатуры. Данный модуль позволял добавлять раскладку клавиатуры для узбекской кирилицы. Для дублирования функция ввода данный симовлов была предусмотрена так же в строке поиска в интерфейсе словаря.

В силу неполной поддержке формата Unicode в ОС этой серии, пришлось дорабатывать содержимое файлов, отвечающих за кодовую страницу. Это было, необходимо для корректного отобра-

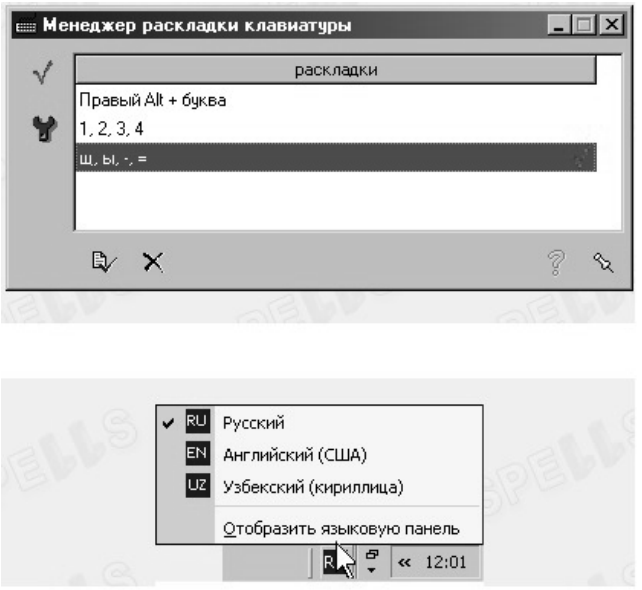


Рис 3. Редактор узбекской клавиатуры

жения вышеуказанных символов в приложениях. В последующих версиях ОС, с полной поддержкой Unicode это иногда приводило к некоторым конфликтам при индексировании некоторых файлов. Данную проблему приходилось так же решать.

4. Транслитератор узбекской кириллицы на узбекскую латиницу

Следующим был разработан модуль транслитерации текста с узбекской кириллицы на латиницу и в обратном направлении. Сами правила замены группы символов не представляют сложности. Основная задача состояла в том, чтобы создать словарь слов, которые не транслитерируются по правилам, а являются исключениями. К примеру «Цирк» транслитерируется как «Sirk», хотя по правилам должно транслитерироваться как «Tsirk». В словарь было добавлено свыше 2000 слов исключений. Подобным словарем обладала только эта программа среди существовавших в то время на рынке.

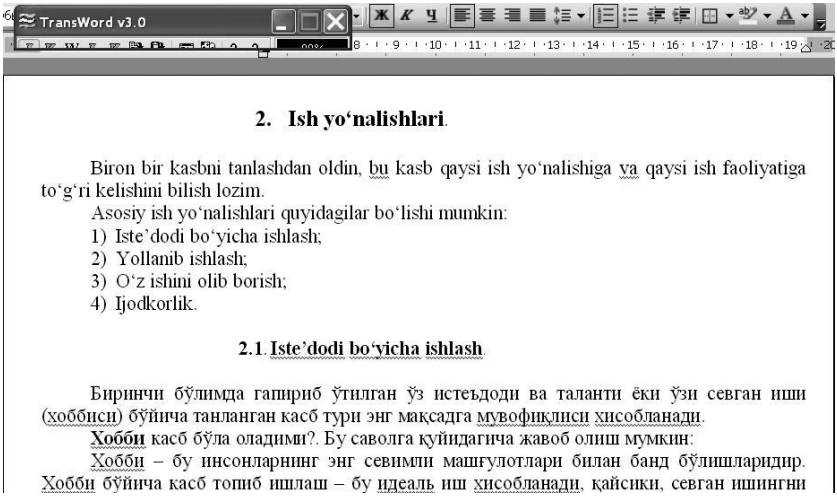


Рис 4. Транслитератор узбекской кириллицы

5. Программа проверки правописания узбекского языка в приложениях Microsoft Office

Далее на основе существующей базы слов узбекского языка было разработано ПО для проверки правописания узбекской кириллицы в приложениях Microsoft Office. Для этого было заключено лицензионное соглашение с компанией Microsoft на использование API приложений данной компании. При создании БД для данного приложения в специальном инструментарии указывали принадлежность каждого слова к той или иной части речи и на основе этого генерировались те или иные группы производных для этих слов. В последующем эти группы производных и служили в качестве эталона при проверке правописания. Компания Microsoft приобрела права на использование подобной базы в релизах своих программ для узбекской латиницы. Это было продиктовано тем, что официальным языком при оформлении документов на территории Республики Узбекистан была признана именно латиница и поддержка кириллицы была признана нецелесообразной.

6. Автоматический переводчик текстов с узбекского языка на русский и в обратном направлении.

На основе созданной базы слов и производных и накопленного опыта было принято решение о создании автоматического переводчика текстов. В течении 2 лет были созданы переводчики текстов сначала с русского языка на узбекский, потом и в обратном направлении. Программы работали по следующему алгоритму. Вначале анализировались слова в исходном предложении. Выделялись основы, по полученным результатам в существующей базе происходил поиск соответствующего перевода и полученные результаты формировались в конечное предложение согласно правил языка, на который осуществлялся перевод. Применение подобного подхода позволило избежать некорректных результатов, которые получается если программа использует прямой метод перевода без учета правил языка, когда исходное слово и полученный перевод остаются в предложении на тех же местах. В инструментарии использовалась единая база для корректоров, работающих в проекте с выдачей статистики по производительности текущему прогрессу.

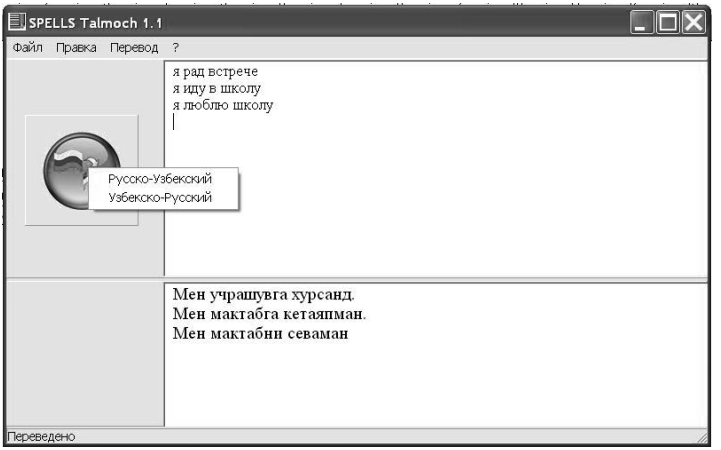


Рис 5. Рус-Узб и Узб-Рус переводчик текстов

Запланированное дополнение в виде возможности переводить текстов на узбекской кириллице и выдачей результатов на ней же не было осуществлено в силу достаточно больших трудозатрат.

При создании продуктов использовались средства разработки C++, Delphi.

Основной штат программистов был 3 человека.

Количество лингвистов и корректоров в разные периоды создания продуктов составлял от 2-до 8 человек. Наибольший рост штата компании был во время работы над автоматическим переводчиком текстов.

Из существующих подобных решений хотелось бы отметить переводной словарь IBORA, компании O‘zbegim Dasturlari. Словарь позволяет получить перевод в трех направлениях-русском, узбекском и английском. Словарные статьи создавались полностью «с нуля», с привлечением лингвистов. Общая база слов составляла около 120 тыс. К сожалению, примерно начиная с 2007 года работа над данным продуктом прекращена.

Так же заслуживает упоминания интернет-ресурс Solver.uz предоставлявший возможность получить перевод в трех направлениях-узбекском, русском и английском. Получить перевод можно было как отдельного слова, так и небольшого текста. В настоящее время работа над данным проектом прекращена. Данные о возможном возобновлении работы ресурса к сожалению, получить не удалось.

Периодически создавались словари для перевода в том или ином направлении, но они не заслуживают внимания в силу маленькой базы слов, что существенно снижает ценность этих продуктов для конечного пользователя.

Заключение

Резюмируя вышесказанное хотелось бы отметить, что в настоящее время, работы над программными продуктами компании «SPELLS» приостановлены, в силу малой востребованности данных продуктов у пользователей и слабой модели монетизации. В настоящее время ведется поиск инвестора для финансирования возобновления дальнейших работ над данными программными продуктами, с возможностью создания на базе существующего ПО онлайн-решений в области перевода как отдельных слов, так и текстов. Проработан проект ресурса для коллективной работы над словарными статьями по принципу Википедии-это позволит существенно ускорит наполнение базы словаря и увеличит качество полученных материалов.

EVALUATION CRITERIA FOR IT TERMINOLOGY (IN THE CASE OF THE MEDIAWIKI INTERFACE)

Nikolai Pavlov

CyberSakha, Yakutsk
halan@yandex.ru

Terms and other special vocabulary words are important for the development of the niche or professional strata in any natural language, particularly for those that come under the pressure of more widely spoken languages. For minority languages, the creation of special terminology for information technology (IT) is of exceptionally high importance because the IT has become a meta-environment for all fields of human activity, so that the vocabulary of IT is rapidly turning from a specialist jargon into a body of words used by an ever higher part of the speakers of any language. Considering this, and also the exceptionally short time frame available for the creation of new language norm, the quality of offered terms is crucial. The criteria collections offered may, as far as possible, objectivize the expert evaluation, and, more importantly, may help the compilers of terminology resources to communicate fast and understand the colleagues successfully and minimize creative tensions.

Развитость лексики того или иного языка в какой-то степени свидетельствует об уровне культуры народа. Если этот тезис, в общем-то, может быть спорен, то значение терминологического аппарата, лексики для специальных целей в современных условиях неоспоримо. Язык, чтобы оставаться функциональным, должен осваивать новые для себя сферы, в том числе инфокоммуникационные технологии, по сути ставшие мета-средой в которой с недавних пор существуют и посредством которой практически полностью стали функционировать все остальные виды деятельности человека.

Поскольку революция в этой области, как и любая другая революция, с исторической точки зрения, происходит стремительно, новая лексика не успевает стать узуальной речевой нормой в общепринятом смысле. Проще говоря, широкие массы народа и профессиональные группы не только не успевают переварить множество вариантов слов специальной лексики, чтобы естественным путем остановиться на наиболее приемлемом из них, но даже познакомиться с ними в полном объеме. В итоге, термин принимается в неизменном виде, нарушающем фонетический строй языка, либо, зачастую, в варианте, не удовлетворяющем часть пользова-

телей компьютерных технологий, что не способствует устойчивости нового термина.

Вследствие этого, еще более возрастает роль составителей терминологических ресурсов (терминологов, лексикографов, специалистов в узких областях знаний), от ответственности и компетентности которых зависит будущее специальной лексики языка, а во многом, судьба самого языка.

Терминотворчеством и фонетическим преобразованием перенимаемых интернациональных слов, в той или иной степени, занимаются многие специалисты. Основными двумя группами являются лингвисты-терминографы и специалисты в узких областях знаний, использующие язык в своей повседневной профессиональной деятельности. Между ними и внутри самих групп могут быть разные подходы к терминообразованию, к оценке специальной лексики, которые обусловлены как уровнем общих лингвистических знаний и уровнем знаний о словообразовательных возможностях, свойственных тому или иному языку, так и "общественно-политическими" причинами.

Эти разные подходы могут обусловить конфликты, потери времени, что в конечном счете приводит к негативным, с точки зрения широты использования языка, последствиям. Объективная оценка терминов могла бы не только поднять уровень дискуссий с категории "нравится/не нравится" до приемлемого аргументационного уровня, но и самим фактом своего существования привести специалистов разной квалификации к единой терминотворческой матрице/парадигме/системе. С другой стороны, объективизация может защитить специалиста от необоснованных нападок со стороны политиканствующих субъектов.

Учитывая это, в 90-х годах XX века терминологическая группа Научно-исследовательского института гуманитарных исследований АН Республики Саха[1] (руководитель группы Оконешников Е.И.) разработала концепцию и принципы терминообразования для языка саха и, основанные на них, критерии оценки терминов. Нужно сказать, что предложенная система оценки является экспертной, и поэтому не может считаться полностью объективной. Вместе с тем, она максимально рационализирует оценку, тем са-

[1] С 2008 г. называется "Институт гуманитарных исследований и проблем малочисленных народов Севера СО РАН (ИГИиПМНС СО РАН)

мым снижая процент ошибок и при использовании достаточным количеством экспертов можно избежать погрешностей, обусловленных личным отношением.

В этой статье хотим показать небольшой опыт применения этих критериев в области ИКТ, в частности на примере локализации интерфейса CMS[1] MediaWiki.

Система управления содержимым (контентом) MediaWiki обеспечивает функционирование таких крупных сайтов, как многоязычная энциклопедия Википедия и библиотека электронных текстов на разных языках Викитека, а также десятки тысяч других сайтов по всему миру. Перевод интерфейса этой системы (локализация на тот или иной естественный язык) производится на специальном сайте TranslateWiki.net волонтерами – носителями языков. По состоянию на июнь 2015 года на этом сайте для 226 активных языков локализовывалось 29 проектов с, в общей сложности, 61 тысячей интерфейсных сообщений. На сайте было зарегистрировано 7,7 тысяч переводчиков. Интерфейсные сообщения или единицы интерфейса (слова, словосочетания, предложения, тексты и аббревиатуры), подлежащие переводу, сгруппированы по важности и функциональному назначению в специальные группы. Единицы интерфейса MediaWiki категоризированы в 5 основных групп, которые насчитывают от 594 до 20 тысяч единиц.

Начало локализации MediaWiki на язык саха датируется апрелем 2007 года, когда было принято решение начинать переводить интерфейс Саха Википедии – раздела Википедии на языке саха (якутском). За время локализации сделано более 12 тысяч правок. В настоящее время актуальных переводов насчитывается 8712 единиц.

Нами проведена пробная выборочная оценка некоторых, часто используемых в интерфейсе MediaWiki, слов.

Система экспертной оценки терминов

Система экспертной оценки, предложенная членом терминологической группы ИГИ Л.А. Афанасьевым-Тэрис, охватывает три уровня. Первый: Вероятность замены термина аналогом на языке.

---

[1] англ. *Content management system* – Система управления содержимым сайта.

Второй: Предварительная оценка устойчивости термина-замены. Третий: Возможность фонетической трансформации (транскрипции) заимствованного слова, с приближением к фонетическому строю языка.

Для удобства восприятия критерии расположены в виде таблиц.

1. Вероятность замены термина аналогом на языке

Оценка вероятности замены термина аналогом на целевом языке производится по 5 критериям (см. табл. 1 с нашими уточнениями).

Таблица 1

Оценка вероятности замены термина аналогом на языке по Л. Афанасьеву-Тэрис (модернизация Н. Павлова)

Критерии оценки	Баллы		
	-1 балл	0 баллов	+1 балл
1. Степень интернациональности термина	Используется в нескольких языках	Используется в нескольких языках, находящихся в тесной связи	Используется только в одном языке (русском)
2. Терминологичность	Слово имеет все признаки термина	Слово не отвечает всем признакам термина	Общеупотребительное слово, описательное, оценочное слово
3. Наличие аналога/протермина в целевом языке	Не существует	–	Есть аналог или аналоги
4. Степень проникновения языка в сферу использования термина (поле)	Язык практически не используется в данной сфере	Язык используется в данной сфере, но специальная лексика еще не устоялась	Язык широко используется в данной сфере
5. Употребляемость термина	Используется только в узком профессиональном кругу	Используется в немногих профессиональных кругах	Используется значимым множеством (большинством) носителей языка

Разработчик не дает количественных оценок, видимо, такая задача им не ставилась. Можно предположить, что она зависит от целей и задач конкретной терминологической работы. Поэтому, мы предложили собственную шкалу оценки результатов:

- если слово, в результате экспертной оценки, набирает нулевой или отрицательный балл – то термин лучше не переводить и попробовать трансформировать его, приближая к фонетическому строю языка;
- если слово набирает 1–2 балла – термин вполне можно перевести, и его устойчивость в качестве термина будет хорошей;
- если же слово набирает свыше 3 баллов, то его национальный (в нашем случае саха) вариант имеет полные возможности быть успешным в качестве термина.

Для примера разберем 8 понятий (слов), применяемых в интерфейсе CMS MediaWiki: ссылка, файл, сохранение, панель, отмена, ОК, электронная почта, бот (см. табл. 2). Для наглядности таблица транспонирована.

Таблица 2

Примеры разбора слов по вероятности подбора аналогов на языке

Слово		Оценка по критериям					
Понятие (рус. / англ.)	Вариант на языке саха	1. Интернациональность	2. Терминологичность	3. Наличие аналога	4. Поле	5. Употребляемость	Итоговый балл
ссылка / link	сигэ	+1	-1	+1	+1	+1	3
файл / file	билэ	-1	-1	+1	+1	+1	1
сохранение / save	бигэргэтии	+1	0	+1	+1	+1	4
панель (инструментов) / panel	хаптал	+1	-1	+1	+1	+1	3
отмена / escape	алҕас	+1	0	+1	+1	+1	4

ОК / OK	сөп	-1	-1	+1	+1	+1	1
электронная почта / e-mail	электрон буоста (хол-буйа)	-1	-1	-1	+1	+1	-1
бот / bot (robot)	оруобат	-1	-1	-1	+1	-1	-2

Как видим, 4 понятия/слова (*ссылка, сохранение, панель* и *отмена*) с точки зрения возможности замены набрали высокий балл и, по мнению эксперта, их саха варианты имеют все возможности для широкого употребления.

Саха варианты двух понятий (*файл* и *ОК*) также имеют неплохой шанс стать нормативными. Для двух других понятий (*e-mail, bot*), имеющих интернациональное выражение, возможно, стоит поискать формы заимствования (транскрипция, транслитерация).

2. Оценка устойчивости термина-замены

Любое слово специальной лексики может быть заимствовано языком или ему может быть найдена замена из лексики самого языка. Замененный термин следует оценить на устойчивость (вероятность принятия тем или иным профессиональным сообществом и широкими массами носителей языка). Оценка проводится по 7 критериям (см. табл 3).

Таблица 3

Критерии оценки устойчивости термина-замены по Л. Афанасьеву-Тэрис (модификация Н. Павлова)

Критерии	Баллы		
	1 балл	0 баллов	-1 балл
1. Соответствие значению	слово-кандидат соответствует значению изначального термина	неполное соответствие	не соответствует

2. Ясность смысла	слово имеет ясный и конкретный смысл	значение слова ясно носителю, но оно размыто либо обозначает родовой признак; можно понять к чему относится	значение слова забыто или по слову нельзя догадаться об его смысле
3. Количество обозначаемых понятий	слово используется во многих значениях		слово используется только в одном смысле
4. Удачность подбора формы слова для термина	удачен, соответствует законам литературного языка	–	не удачен
5. Терминологичность, символичность	сильный термин	обиходное слово	описательный перевод
6. Соответствие / близость к первоначальному смыслу прототермина (обычно на латыни или древнегреческом)	близок	–	далек
7. Количество вариантов перевода	существует единственный вариант	есть несколько вариантов перевода, но один из них существенно превалирует	в каждом терминологическом поле свой вариант перевода

Автор методики здесь также не дает количественных оценок, поэтому мы предлагаем свой вариант шкалы оценок:

- Если слово набирает нулевой или отрицательный балл – вариант перевода требует доработки;

Таблица 4

Примеры разбора термина-замены

Слово	Оценка по критериям							
	1. Соответствие значению	2. Ясность смысла	3. Коли-чество обозначаемых словом понятий	4. Удачность подбора формы	5. Терминологичность/ символичность	6. Соответствие изначальному смыслу термина	7. Количество вариантов	**Итоговый балл**
	соответствует: +1 не соответствует: -1 промежуточное значение: 0	ясный: +1 непонятный: -1 промежуточная оценка: 0	обозначает одно понятие: +1 обозначает несколько понятий: -1 Промежуточная оценка: 0	удачный: +1 неудачный -1 промежуточный: 0	сильный термин: +1 описательный перевод: -1 Обиходное слово: 0	соответствует +1 не соответствует: -1 частичное соответствие: 0	единственный +1 несколько -1 несколько, но один вариант превалирует: 0	
сигэ	+1	+1	+1	+1	+1	+1	+1	**7**
билэ	0	-1	0	+1	+1	+1	+1	**3**
бигэргэтии	+1	+1	0	+1	+1	+1	+1	**6**
хаптал	+1	0	-1	+1	+1	+1	0	**3**
алҕас	+1	+1	+1	+1	+1	+1	0	**6**
сөп	+1	+1	+1	+1	+1	+1	+1	**7**
тиһик	0	0	0	+1	+1	+1	0	**3**
уҥа баттам	+1	0	+1	+1	+1	+1	+1	**6**
кыттаччы	+1	+1	-1	+1	0	0	0	**2**

- если слово набирает от одного до четырех баллов – вариант приемлемого качества;
- если слово набирает 5–7 баллов, то перевод хорошего качества, можно предположить, что при его использовании даже неквалифицированными пользователями, те не будут иметь затруднений.

Рассмотрим это на примере слов, использованных в интерфейсе CMS MediaWiki: сигэ (ссылка), билэ (файл), бигэргэтии (сохранение), хаптал (панель), алҕас (ошибка), сөп (ОК), тиһик (система), уҥа баттам (правая кнопка мыши), кыттааччы (участник).

Все представленные примеры переводов, судя по экспертным оценкам, удовлетворительного или хорошего качества, это позволяет предположить высокую вероятность приживаемости данных терминов.

3. Возможность фонетической трансформации заимствованного термина

Заимствованный термин, как правило, подлежит транскрибированию. В поздний советский период существовало правило правописания, требующее оставлять заимствованные после Октябрьской революции русские слова (или интернациональные слова, вошедшие в язык посредством русского) без изменения. В 90-е гг. это правило было отменено, но инерция такова, что даже спустя десятилетия, оно продолжает иметь своих последователей по большей части в нелингвистической среде. Существуют и объективные причины, мешающие транскрибированию согласно фонетическим законам языка. Л.А. Афанасьев-Тэрис предложил 7 критериев, с помощью которых лексикографы могут принять решение (см. табл. 5).

Таблица 5

Критерии возможности трансформации заимствованного термина по Л.Афанасьеву-Тэрис (модификация Н. Павлова)

Критерии	Баллы		
	+1	0	-1
1. Длительность использования заимствованного слова	заимствование используется в языке долгое время		заимствовано недавно

2. Широта использования	широко	в нескольких профессиональных сферах	в узкопрофессиональной сфере
3. Частота использования	используется часто	в некоторых сферах используется часто	малоупотребимый
4. Степень освоения языком поля	язык освоил поле	промежуточное положение	язык не освоил поле
5. Практика использования	фонетизированный вариант использовался ранее	фонетизированный вариант используется в устной речи, но в письме не встречается	об использовании фонетизированного варианта в устной речи неизвестно, в письме не встречается
6. Трудность произношения/ произнесения	заимствование удобно для произношения носителем языка	промежуточный вариант (есть звуки нехарактерные для языка, но закона сингармонизма не нарушают)	сложен для произношения носителем языка
7. Новизна понятия (неологизм)	придумано недавно	в некоторых профессиональных сферах укоренилось	давно существующее, укорененное понятие

Предлагаем следующую шкалу оценки:

- Если термин набирает баллы от 0 и выше, то вероятность принятия фонетизированного варианта высока;
- если набирает отрицательный балл – высока вероятность использования в русской форме.

Рассмотрим три пары терминов также используемые в вики-интерфейсе: электронная почта – электрон буоста, бот – оруобат, индексация – ииндэкистээһин.

Таблица 6

Пример экспертной оценки заимствованного термина (сами критерии для экономии места не приведены, их номера соответствуют номерам критериев из табл. 5)

Слово		Баллы							
Русский вариант	Предлагаемое написание в саха тексте	1. давно +1 новое -1 промежуточное 0	2. широко +1 узко -1 промежуточное 0	3. часто +1 редко -1 промежуточное 0	4. освоил +1 не освоил -1 промежуточное 0	5. есть +1 нет -1 промежуточное 0	6. удобно +1 неудобно -1 промежуточное 0	7. новое +1 существующее -1 Промежуточное 0	**Заключительный балл**
электронная почта/ e-mail	электрон буоста	0	+1	+1	0	-1	0	+1	**2**
бот / bot	оруобат	+1	-1	0	0	+1	+1	+1	**3**
индексация / index	ииндэкистээһин	0	-1	-1	0	-1	0	+1	**-2**

Судя по приведенной здесь экспертной оценке, два термина могут быть легко использованы в фонетизированном варианте, для третьего термина (*индексация/index*) высока вероятность использования в русском варианте. В этом случае можно, в качестве альтернативы, порекомендовать поискать переводной аналог или, если позволяет контекст (графический интерфейс), то сделать описательный перевод.

В рамках данной статьи нами сделана попытка адаптировать ранее предложенный группой терминологов ИГИ АН РС(Я) принцип оценки терминов для языка саха под нужды IT сферы. Применение экспертных критериев Л.А. Афанасьева-Тэрис в ИКТ возможно и может повысить качество терминологической работы, снизить вероятность конфликтов и защитить специалиста, предлагающего термины, в период формирования и освоения термина в той или иной области профессиональной деятельности от необоснованных нападок.

ЛИТЕРАТУРА

1. Афанасьев Л.А. Саха тылын сайдар туруга. С. 22–39 // Сахалыы тиэрмини оҥоруу (Вопросы якутской терминологии, Сборник научных трудов). Дьокуускай, 1995.

2. Афанасьев Л.А. Тиэрмини оҥорор үлэ толору көрүҥэ. С. 39–47 // Сахалыы тиэрмини оҥоруу (Вопросы якутской терминологии, Сборник научных трудов). Дьокуускай, 1995.

3. Афанасьев Л.А. Тиэрминнэри сыаналыыр мэктиэлэр. С. 47–63 // Сахалыы тиэрмини оҥоруу (Вопросы якутской терминологии, Сборник научных трудов). Дьокуускай, 1995.

4. Слепцов П.А. Тиэрмин сайдыытын сүрүн туһаайыылара. С. 64–70 // Сахалыы тиэрмини оҥоруу (Вопросы якутской терминологии, Сборник научных трудов). Дьокуускай, 1995.

5. Быганова В.И. Саха литературнай тылыгар тиэрминнэри оҥоруу уонна бэрээдэктээһин сүрүн тосхоллоро. С. 96–103 // Сахалыы тиэрмини оҥоруу (Вопросы якутской терминологии, Сборник научных трудов). Дьокуускай, 1995

6. Оконешников Е.И. Лингвистические аспекты терминологии языка саха. Якутск, Издательство СО РАН, Якутский филиал, 2004. – 196 с.

7. Сулейманов Д.Ш., Галимянов А.Ф. Система татарских терминов в компьютерных технологиях и информатике. // Труды Казанской школы по компьютерной и когнитивной лингвистике. Выпуск 15. Академия наук РТ, Казань, 2012. С. 61–69.

8. Финикова И.В. Новый подход к пониманию природы термина. Вестник МГИМО-университета. Выпуск № 3 / 2012. С. 177–181.

9. Заглавная страница сайта перевода программ с открытым исходным кодом TranslateWiki.net https://translatewiki.net/

10. Миньяр-Белоручева А.П. О сходстве и различии терминов и номенклатурных образований. Вестник Чувашского университета. Выпуск № 4 / 2014. С. 166–170.

11. Павлов Н.Н. Көмпүүтэр бырагырааматын уонна ситим-сир алтыһаанын сахалыы тылбааһа (Саха интерфейс компьютерных программ и сайтов, автореферат магистерской диссертации), Якутск, СВФУ, 25.06.2015. 20 с.

PROJECT OF ELECTRONIC ETHNO-LINGUISTIC TATAR DICTIONARY[1]

Farid Salimov, Rustem Salimov

Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences,
Kazan Federal University
Kazan, Tatarstan, Russia

We analyse the first step of electronic ethno-linguistic resource creation (ethnoling.antat.ru), built on the basis of ethno-linguistic expeditions of the Institute of Language, Literature and Art (ILLA) of Academy of Sciences of the Republic of Tatarstan. More than 20 books and 300 scientific articles were published on the basis of systematization of materials collected by ethno-linguistic staff of ILLA in different years. Materials were collected in respect of ethno-cultural archaic dialect zones of Siberia, the Urals region, the Middle and Lower Volga region, densely inhabited by Tatar population.

The purpose of this project is to create an electronic resource which includes information in a structured way, extracted from major publications, based on the results of ethno-linguistic expeditions of ILLA. As a result of completion of the project electronic resource should be created and posted in the Internet. It should contain the terminology (ethno-linguistic) dictionaries with large amounts of live specimens of the Tatar language, collected in the expeditions. In addition, it is expected to bind created resources with electronic atlas of Tatar dialects.

Введение

В современном мире наблюдается повышенный интерес к языку духовной народной культуры. Материалы живой диалектной речи и народной культурной традиции являются ценнейшим источником сведений о духовной культуре народа.

Начиная с 60-х годов XX столетия в институте языка, литературы и искусства им. Г. Ибрагимова Академии наук Республики Татарстан (ИЯЛИ АН РТ) проводится работа по сбору этнолингвистического материала по диалектам и говорам татар, проживающих в Республике Татарстан и других регионах России. Сбор этих данных проходил параллельно со сбором информационных материалов для атласа татарских народных говоров [Атлас татар-

[1] Работа выполнена при финансовой поддержке РГНФ в рамках проекта Создание электронного ресурса по этнолингвистическим (диалектно-фольклорным) материалам татарского языка (проект № 14-04-12024)

ских народных говоров Среднего Поволжья и Приуралья, 1989]. И хотя географические зоны полевых экспедиций во многом пересекались, программы научных исследований были различными, в результате лексический материал, собранный в рамках двух направлений отличался и дополнял друг друга. На основе анализа и систематизации собранных этнолингвистических материалов сотрудниками ИЯЛИ в различные годы было опубликовано более 20 монографий и около 300 научных статей. Материалы были собраны в архаически этнокультурном отношении диалектных зонах Сибири, регионах Урала, Среднего и Нижнего Поволжья, где компактно проживает татарское население.

Электронный словарь

В 2011–2012 годах при финансовой поддержке РГНФ (проект №11-04-12020в) был создан электронный атлас татарских народных говоров (atlas.antat.ru), в котором на 215 картах отражены основные признаки диалектного различия для татарского языка, закономерности и характер распространения диалектных явлений в 28 регионах Российской Федерации [Салимов Ф.И., 2012]. Издание электронного атласа представляло первый опыт по интерпретации и представлению обширнейшего материала, изданного ИЯЛИ АН РТ в рамках печатного варианта атласа. Имеющиеся материалы были значительно дополнены дополнительными данными, собранными после 1989 года в сибирских регионах, а также библиотекой стандартных параметризованных запросов, позволяющих проводить многоаспектный анализ имеющегося материала с привязкой языковых явлений к географическим координатам.

Целью настоящего проекта является создание электронного ресурса, включающего в себя в структурированном виде информацию, извлеченную из основных публикаций, изданных по результатам этнолингвистических экспедиций ИЯЛИ АН РТ. После завершения проекта, в сети Интернет должен появиться электронный ресурс, в котором будут представлены терминологические (этнолингвистические) словари с большим объемом образцов живой татарской речи, собранных в полевых условиях. Кроме того предполагается привязать создаваемый ресурс к картам электронного атласа татарских говоров.

Основным источником для создания словаря были выбраны фундаментальные труды Ф.С. Баязитовой [Баязитова 2011, 2012], которые освещая в зависимости от разрабатываемой тематики различную диалектную лексику, имеют сходную конструкцию, позволяющую применять к анализу имеющегося материала похожие инструменты.

В этой статье описан опыт анализа научной монографии Ф.С. Баязитовой «Лексика татарской свадьбы (в контексте диалектных и фольклорных текстов)» /«ТУЙ ЙОЛАЛАРЫ ЛЕКСИКАСЫ (җирле сөйләш һәм фольклор текстлары яссылыгында)»/, изданной в 2011 году в Казани. В этой научной монографии на примерах различных образцов диалектных текстов исследуются исторические корни возникновения различных свадебных терминов, приводятся многочисленные примеры использования этих терминов в народной речи.

Монография представляет собой объемное научное исследование (около 960 страниц текста), которое включает большое количество иллюстративного материала, представляющего собой диалектные тексты в упрощенной транскрипции, записанные и расшифрованные с магнитных лент. Все приводимые примеры разнесены по темам, связанным с обрядовой свадебной лексикой, и сопровождаются семантическими толкованиями отдельных терминов. Диалектные лексические термины в большей своей части содержат ссылку на диалекты и говоры, к которых они встречаются.

Построение этнолингвистической базы данных предполагало выполнение определенной последовательности шагов и включало в себя следующие этапы:

1. Сканирование печатного экземпляра книги с целью создания ее электронного образа;

2. Анализ и систематизация имеющегося в книге материала;

3. Создание программы сегментации линейного текста с выделением набора необходимых фрагментов, вносимых в электронную базу данных; Автоматическая сегментация содержимого книги с последующей ручной проверкой и корректировкой выделенных сегментов;

4. Проектирование и создание базы данных, заполнение таблиц полученными данными;

5. Построение запросов к базе данных;

6. Создание клиентской части программы, размещение тестовой версии программы в Интернет.

1. Содержание книги Ф.С. Баязитовой представляет собой богатейший этнолингвистический источник, в котором с позиции различных диалектов татарского языка анализируется такая важная тематика, как свадебный обряд. Представленный в книге материал имеет ценность как в лингвистическом, так и этнографическом аспектах. Особую ценность представляет множество примеров практического использования терминов в диалектной речи татар. Поскольку у разработчиков в наличии имелся только печатный экземпляр книги, было проведено сканирование печатного материала с целью его перевода его в электронный образ с сохранением особенностей форматирования исходного текста. Эта работа с последующим ручным исправлением ошибок результатов сканирования потребовала достаточно больших усилий в виду большого объема исходного текста.

2. Книга Ф.С. Баязитовой состоит из многочисленных тематических групп (свадебные ритуалы, свадебные персонажи, сваты и сватовство, свадебная пища, свадебная одежда, и пр.), в каждом из которых описывается и систематизируется определенный набор свадебных терминов, относящийся к соответствующему разделу. В тексте книги перемешаны как сами термины, примеры их употребления в различных диалектах и говорах, семантическое толкование термина и другая информация. Основная задача анализа состояла в выявлении системы признаков, характеризующих различные фрагменты текста: термины, ссылки на диалекты и говоры, в которых они употребляются, их семантическое описание, а также тексты примеров, которые служат для иллюстрации употребления соответствующих терминов в различных диалектах и говорах. Такой анализ представлял достаточно трудную задачу. Поскольку, изначально, книга не предполагала автоматическую обработку имеющегося материала, не всегда выдерживался единый язык разметки терминов, многие одинаковые в семантическом отношении фрагменты были оформлены различным образом, в приводимых словоформах не везде проставлены ссылки на диалекты и говоры, а имеющиеся ссылки – часто относились к различным по объему и содержанию текстовым фрагментам. В силу описанных причин,

в определенных случаях, было довольно трудно автоматически, с помощью программы сегментации, определить на какой длины текстовый фрагмент распространяется приведенное значение параметра. Для выделения границ сегментов применялись различные подходы, в частности, использование форматов представления определенных фрагментов предложений в печатном варианте текста книги, анализ информации из некоторой окрестности анализируемого термина, построения словарей, содержащих наборы ключевых слов.

3. Была написана программа сегментации текста с выделением ключевых терминов, описания их семантики, указания диалекта или говора в которым встречается соответствующий термин, выделения примеров употребления термина в различных диалектах. Результаты работы программы сегментации представляются в виде набора специальных таблиц в формате текстового процессора WORD. При сегментации также учитывались расшифровки диалектизмов, которые встречаются в тексте книги и вынесены из текста в виде сносок. Эти диалектизмы составили отдельный словарь. К сожалению, вариативность представления информации в различных фрагментах текста оказалась достаточно большой, и несмотря на предпринятые усилия, не удалось полностью автоматизировать процесс сегментации. Поэтому,полученный на выходе программы размеченный текст дважды подвергался ручной обработке: сначала текст проверялся на ошибки определения границ выделенных фрагментов, далее откорректированный текст проверялся лингвистами на предмет его дополнения недостающими параметрами (в основном, указанием на диалекты и говоры, где употребляется тот или другой термин).

4. На основе отобранного материала был создан терминологический словарь диалектизмов (словоформ и словосочетаний). В состав словаря дополнительно была включена информация по транскрипции диалектизмов, а также перевод семантики включенных в словарь терминов на русский язык. При создании базы данных облегченная транскрипция, используемая Ф.С. Баязитовой была признана недостаточной и для каждого термина была предложена транслитерация терминов на письменные формы татарского языка с использованием символов Международного фонетического алфавита (МФА) [У.Ш.Байчура]. При этом преследовалась цель

использования терминов словаря в диалектических корпусах. К настоящему времени общий объем словаря составляет около 3500 словоформ и словосочетаний.

В базу данных была включена обширная коллекция диалектных текстов, собранных в полевых экспедициях. Элементы базового словаря и примеры связаны между собой отношением один ко многим, которое позволяет по заданному термину найти набор примеров по употреблению данного термина в различных диалектах. Была разработана специальная программа согласованной загрузки данных из обработанных текстовых фрагментов. Дополнительно, данные базового словаря были индексированы специальными пометками, что позволило строить запросы по набору словоформ, относящихся к отдельным разделам свадебной терминологии. В состав базы данных также были включены дополнительные словари по диалектизмам татарского языка, которые непосредственно не относятся к свадебной тематике, но встречаются в тексте книги и расшифровываются в виде сносок. Дополнительно в базу данных включена информация по населенным пунктам, где собиралась информация, по информантам, по научным источникам, которые были использованы при создании книги. При анализе списка населенных пунктов добавлена информация по диалектам и говорам, характеризующим данный населенный пункт, эта информация сравнивалась с подобной информацией для населенных пунктов, описанных в составе атласа татарских говоров. Следует отметить, что из более чем 200 населенных пунктов, где собиралась информация по свадебной лексике, только лишь половина описана в составе атласа. Это объясняется тем, что информация по этнолигвистическим материалам в основном собиралась в отдаленных, малонаселенных селах, в которых в силу их удаленности от «цивилизации», сохранились сведения о первобытном языке и культуре. Однако это обстоятельство создало определенные сложности при определении географических координат соответствующих населенных пунктов. Со времени проведения экспедиций, многие населенные пункты прекратили свое существование, и было довольно трудно найти по ним необходимую информацию.

5. Во процессе аботы над проектом создана библиотека типовых запросов по следующим направлениям:

• Запросы, которые по выбранной словоформе (словосочетанию) и диалекту (говору) показывают на экране наборы образцов текстов, иллюстрирующих использование этой словоформы (словосочетания) на примерах живой народной речи.

• Построения списка словоформ, связанных с некоторой темой свадебного обряда.

• Запросы по формированию карт, которые показывают географию распространения соответствующей словоформы. Поскольку в книжной версии не указан конкретный список населенных пунктов, связанных с распространением того или иного термина, при построении карт за основу было взято разбиение населенных пунктов по диалектам и говорам, зафиксированное в электронном атласе народных говоров. Такое представление, не являясь абсолютно точным, все же дает определенную картину о географическом распространении лексических терминов, использованных при написании книги.

• В настоящее время производится индексация базового словаря, которая позволит строить запросы для выявления терминов-синонимов, употребляемых в различных диалектах.

6. Программа создается как интернет-ресурс. Тестовый вариант программы доступен по адресу ethnoling.antat.ru. Основную часть программы составляет базовый словарь терминов, который позволяет по заданному термину строить множество примеров, в которых встречается употребление этого термина. Визуализация содержимого словаря в программе может производиться в различных режимах: в алфавитном порядке, в тематическом плане. Выбор режима просмотра определяется конечным пользователем. В программе реализовано два режима работы: режим пользователя, и режим администратора системы. Режим администратора необходим для модификации структуры базы данных и заполнения ее содержимого новыми данными.

Заключение

Представленный в базе данных материал может служить ресурсом для создания диалектологических подкорпусов татарского языка, этнолингвистических словарей, в частности построения словаря диалектизмов.

В будущем, предполагается пополнение словаря новыми словоформами, которые позволят точнее дифференцировать диалекты татарского языка, полнее описывать их особенности. Отметим, что в словарь вошло много словоформ и словосочетаний, которые отсутствуют в Большом диалектологическом словаре татарского языка 2009 года издания.

ЛИТЕРАТУРА

1. У.Ш.Байчура «Звуковой строй татарского языка», изд-во КГУ, 1959 г.

2. Атлас татарских народных говоров Среднего Поволжья и Приуралья / науч. редакторы: Н. Б. Бурганова, Л. Т. Махмутова; Составители: Н. Б. Бурганова, Л. Т. Махмутова (I т.), Ф. С. Баязитова, Д. Б. Рамазанова, З. Р. Садыкова, Т. Х. Хайрутдинова (2 т.). – Казань, 1989. Прил.: Комментарии к Атласу. – Казань, 1989. – 300 с.

3. Баязитова Ф.С. (2011) Туй йолалары лексикасы (җирле сөйләш һәм фольклор текстлары яссылыгында). – Казан: Паравитта, 2011. – 976 б.

4. Баязитова Ф.С. (2012) Халык традицияләре лексикасы: бишек туе (йола һәм фольклор текстлары яссылыгында). – Казан: Алма-Лит, – 331 б.

5 Салимов Ф.И., Рамазанова Д.Б., Пилюгин А.Г., Салимов Р.Ф. (2012) Электронная версия атласа татарских народных говоров// Вестник татарского государственно-гуманитарного педагогического университета, Кзань, с. 205–210.

6. Салимов Ф.И. Пилюгин Г.А. Ершов С.А. (2012) Электронный атлас татарских народных говоров как инструмент исследования – Труды татарской школы по компьютерной и когнитивной лингвистике, TEL-2012, ФЭН, АН РТ, Казань, с. 48–54.

AUTOMATING THE BILINGUAL DICTIONARIES CREATING PROCESS BASED ON THE CONVERTATION. FROM THE EXPERIENCE OF THE MARI-RUSSIAN DICTIONARY CREATION

Andrey Chemyshev[1], Andrey Boltachev[2]

Moscow, Russia
[1]chemyshev.andrey@gmail.com, [2]andrewboltachev@gmail.com

For many of the titular languages of the former USSR, including the Turkic languages, there are good dictionaries in one direction – from the national language to Russian. Dictionaries in the other direction are either non-existent, or are of significantly lower quality. Many of them have been compiled using old-fashioned methods – by using a paper card index, or by fully manual typesetting (e.g. through use of WYSIWYG-editors, such as Microsoft Word or other programs). This is a more labour-intensive technology which increases the volume of routine work and has a negative influence on the quality, and does not contribute to the development of creative solutions. The essence of automation is the application of a finite amount of effort to the handling of any volume of data. The compilation of dictionaries involves very large volumes, and the uniformity of the data is also high (word articles are all similarly structured). This article describes the technology for creating bilingual dictionaries, using a Mari-Russian dictionary as an example. This technology can also be used for Turkic languages. The article describes how 'reversal' programs can greatly simplify the creation of dictionaries in the other direction.

1. Общее представление

Для многих титульных языков РФ в постсоветские десятилетия изданы полные академические словари, например, марийско-русский (10 томов), удмуртско-русский и т. д. При этом полными русско-национальными словарями в настоящее время мало кто может похвастаться, а потому в обиходе российских автономий продолжают использовать словари 60–70-х годов прошлого века, которые далеко не отвечают требованиям сегодняшнего дня.

Но и новейшие издания, составленные традиционными «ручными» способами, весьма далеки от совершенства. Так, анализ большого Русско-коми словаря (2003 год, 50 тыс. словарных статей) показал, что словарь по качеству резко уступает Коми-русскому словарю 2000 года: в нём встречается огромное количество оши-

бок, неточностей, иллюстративные примеры взяты «с потолка» некоторые термины некорректно переведены. Примеры на сочетаемость слов также минимальны. Представляется, что если бы авторы при создании данного словаря активно опирались, по крайней мере, на базу Коми-русского словаря, то многих недостатков можно было бы избежать.

Мы считаем, что в больших русско-национальных словарях словарная статья должна в обязательном порядке содержать, помимо прочего, набор словосочетаний, иллюстрирующих употребление русского слова в контекстах с эквивалентами этих словосочетаний на национальном языке. Примеры словосочетаний для начала можно брать из национально-русских словарей, составление которых обычно проходило с опорой на картотеку, отражающую реальное словоупотребление в литературном языке.

2. Исходные данные

В данном докладе мы опишем алгоритм создания русско-марийского словаря на основе марийско-русского словаря. При этом укажем, что данный алгоритм применим для любого финно-угорского, тюркского и других языков.

Базы многих электронных словарей (татарско-русского, башкирско-русского, тувинско-русского и др.), в основном, находятся в формате DSL – этот формат разработан ABBYY и может использоваться в некоторых open source программах, например, в оболочке для электронных словарей GoldenDict.

Для обработки словарей данный формат не очень удобный, поэтому словари «перегоняем» (конвертируем) в базу данных. С обычной базой данных возникает проблема – структура словарных статей даже в одном словаре неоднородная, их великое множество. Поэтому было решено выбрать документо-ориентированную систему управления базами данных с открытым исходным кодом, не требующую описания схемы таблиц. Одним из таких систем является MongoDB.

Для создания русско-марийского словаря (или русско-удмуртского, русско-тувинского, русско-хакасского) необходимо решить сколько будет словарных статей. Прежде всего для будущих словарей необходимы словники. Нами подготовлены несколько ва-

риантов словников: на 50 тысяч слов и 12,5 тысяч. Обычно такое количество словарных статей в так называемых «больших» и «малых» словарях.

Фрагмент словника выглядит так:

…

апельсиновый
апельси́новый
-ая, -ое
то же, что: апельсинный

аплодировать
аплоди́ровать
несов. кому-чему

…

Здесь приведено слово, далее то же слово с ударением и некоторые пометы, такие как часть речи и другие.

3. Используемые готовые материалы и разработки

Берётся для работы база марийско-русского словаря на основе 10-томника: «Марий йылме мутер» / «Словарь марийского языка» в 10 томах. Йошкар-Ола: Марийское книжное издательство, 1990– 2005[1]. В эту базу добавляются другие терминологические и иные словари, например: Словарь лингвистических терминов марийского языка (Марий йылмышанче терминологий мутер) – Сомбатхей, 2005[?], Словарь цветообозначающих слов современного марийского языка (Кызытсе марий йылмысе тӱс лӱм-влак) – Сомбатхей, 2008[2]. В связи с тем, что в марийском словаре в 2011 году изменилась орфография, вся марийская часть базы приводится в соответствии с Марий орфографий мутер / Йылмым, литературым да историйым научнын шымлыше В.М. Васильев лӱмеш марий институт. – Йошкар-Ола, 2011[3]. При этом исключается горномарийская часть 10-томника. Для выявления слов, которые была записаны в старой орфографии используется написанный нами скрипт на Python и библиотека HunSpell на C++.

В программе по «перевёртыванию» словаря используется морфология русского языка. Из открытых источников взят русский HunSpell-словарь, произведена его «ё-фикация», то есть проставлена буква «ё» в словах и парадигмах. В перспективе могут быть

использованы и другие морфоанализаторы русского языка, например, на основе XFST-HFST.

4. Алгоритм работы

Программа сначала выводит переводы терминов, а затем – иллюстративные примеры:

…

апельсиновый

апельси́новый

-ая, -ое

то же, что: апельсинный

апельсин; чевер-тулешке

апельсиновое дерево апельсин пушеҥге

апельсиновая корка апельсин шӱм

аплодировать

аплоди́ровать

несов. кому-чему

соваш; совкалаш; кыраш; лупшаш; совым кыраш; совкален налаш

аплодировать в такт такт почеш совым кыраш

аплодируют изо всех сил уло кертмын совым кырат

проводили с аплодисментами кид совен ужатышт

дружно аплодировать рӱжге совкалаш

…

Далее результаты работы скрипта «заливаются» на веб-сервис, где доступ для онлайн-правки (редактирования) даётся филологам, которые участвуют в проекте. Все остальные могут только просматривать и предлагать свои варианты.

Для каждой словарной статьи пометы, переводы, гиперссылки и др. располагаются в отдельных полях. Для удобства восприятия иллюстративные примеры на марийском и русском языках написаны разными цветами. Сделана возможность оперативного добавления новых полей. Условные сокращения предлагаются из заранее подготовленного списка.

В онлайн-сервисе на основе HunSpell реализован также спеллчекер марийского и русского языков, позволяющий значительно облегчить работу филологов-корректоров – слова в полях перевода,

иллюстративных примеров, идиоматических выражений, фразеологических сочетаний, которые «не понимаются» спелл-чекером (это или ошибка, или данного слова нет в HunSpell-словаре) выделяются разными цветами.

Скрипт не идеальный, он допускает ошибки: не понимает омонимы. По-этому ручное редактирование словаря необходимо.

Если иллюстративных примеров недостаточно или их не оказалось в исходной базе, то для пополнения словаря можно воспользоваться корпус-менеджером, например AntConc и корпусом языка первого порядка, т. е. «голым текстом» без помет и грамматических тегов. Использование корпус-менеджера позволит создать более качественный словарь, так как в качестве примеров будут использованы словосочетания, реально употребляющиеся, например, в художественной литературе, СМИ и т. д. Но работа с корпус-менеджером, это уже отдельная тема.

Следует отметить, что не смотря на автоматизацию процесса, интеграция нового материала в уже существующую словарную статью (её заготовку) и редактирование самой статьи будет довольно трудоёмким процессом, но для квалифицированного филолога это не является невыполнимым. По крайней мере данная работа значительно проще, чем традиционные операции лексиколога с бумажной картотекой. Время интеграции и редактирования зависит от количества принимающих участие в работе филологов, их мотивации и квалификации. В оптимальном случае можно решить вопрос в пределах одного года. Как пример можно указать на тот факт, что 1/10 часть русско-коми словаря была довольно качественно обработана одной из участниц проекта за 1,5 месяца, при том что для коми речь идёт об интеграции нового иллюстративного материала в уже существующий словарь.

Специфика работы по редактированию аналогичного словаря детально описана на сайте: http://wiki.fu-lab.ru/index.php/Русско-коми_электронный_словарь

После окончания редактирования русско-марийский словарь конвертируется в несколько форматов: в формат для издания словаря на бумажных носителях и в несколько форматов для офлайн и онлайн электронных словарей: например, в формат DSL – для работы с оболочкой для электронных словарей

GoldenDict и в XML – для работы с CMS GlossWord.

5. Выводы

Таким образом, опыт создания русско-марийского словаря на основе конвертации марийско-русского может быть использован при создании русско-удмуртского и любых русско-тюркских словарей. Тем более в данной технологии заложена возможность оптимизации существующих словарей путём использования иллюстративных примеров самого словаря, расположенных в разных статьях.

СПИСОК ЛИТЕРАТУРЫ

[1] «Марий йылме мутер» / «Словарь марийского языка» в 10 томах. Йошкар-Ола: Марийское книжное издательство, 1990–2005.

[2] Словарь цветообозначающих слов современного марийского языка (Кызытсе марий йылмысе тӱс лӱм-влак). Сомбатхей, 2008.

[3] Йылмым литературым да историйым научнын шымлыше В.М. Васильев лӱмеш марий институт. Марий орфографий мутер. Йошкар-Ола, 2011.

[4] Дмитриев С.Д. Русско-марийский словарь: для школьников. Йошкар-Ола: Марийское книжное издательство, 2013.

SECTION 3 | **ELECTRONIC TEXT CORPORA**

CONSTRUCTION OF UYGHUR INITIAL PARAPHRASE CORPUS

Kahaerjiang Abiderexiti, Maihemuti Maimaiti, Aishan Wumaier, Tuergen Yibulayin[1]

School of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang, 830046, China
Xinjiang Laboratory of Multi-Language information Technology, Urumqi, Xinjiang, 830046, China
[1]turgun@xju.edu.cn

Paraphrases are approximate meaning of the same content including words, phrases and even sentences. Although there is much works done about paraphrase in terms of constructing corpus and study at different levels of language in other languages, Uyghur paraphrases corpus has not been constructed and not been yet studied according to our best knowledge. In this paper we have studied Uyghur paraphrases and have constructed initial paraphrase corpus. First we define our scope of Uyghur paraphrases and then we design and implement Uyghur paraphrase annotation tool. Finally Uyghur native speakers annotate and retrieve paraphrases using annotation tools. The overall size of corpus 5785 pairs, including various type of paraphrases obtained by article annotation and rewritten paraphrases. Paraphrases obtained by article annotation includes 572 lexical paraphrases, 244 phrasal paraphrases, 114 sentential paraphrases, 55 terms and it's definitions. Rewritten paraphrases are 4800 pairs, among them 2400 pairs are consist of paraphrases each other, 2400 pairs are not. These works are now under progress to extend size of corpus.

1. Introduction

Uyghur language belongs to Turkic language family, and mainly used in Xinjiang Uyghur Autonomous Region (XUAR) in China. This

area is ethnic group living area, with the total population of around 23 million people, the minority population account for about 60%. Presently in XUAR, Uyghur language is one of the official language as well as Mandarin (Chinese). Although Uyghur language considered minority language in China, It considered fifths big language among Turkic languages, so the level of Uyghur information processing not only has a crucial role in dissemination and development of culture and technologies in this region but also development of Turkic languages.

This paper reviews various definitions and corpus of paraphrases in English, Turkish, and put forwards adopted four type of Uyghur paraphrase. Then introduce paraphrase annotation tools. Finally describes the process of constructing Uyghur initial paraphrase corpus.

2. Previous work

Paraphrases can be comprehended as different expressions of the same meaning. For a wide variety of natural language processing application paraphrases are very useful. However, "There is a precise and commonly accepted definition of paraphrasing does not exist. From the perspective of linguistics and computational linguistics, the definition of "approximate sameness of meaning" is generally assumed. "(Vila et al., 2014). The Paraphrases may occur at several levels such as lexical paraphrases, phrasal paraphrases and sentential paraphrases (Bhagat and Hovy, 2013; Madnani and Dorr, 2010). Paraphrasing methods perform recognition, generation and extraction of paraphrase pairs (Androutsopoulos and Malakasiotis, 2010).

There are several paraphrase corpora. MSR corpus contains 5801 English sentence pairs that obtained by unsupervised technique from thousands of English online news sources (Dolan et al., 2004). Human annotated paraphrase corpus contains 900 sentences pair and 300 of which are doubly annotated (Cohn et al., 2008). PAN plagiarism corpus 2010 contains 64558 artificial and 4000 simulated plagiarism cases considered to be the first a large and high diversity of artificial and simulated plagiarism cases (Potthast et al., 2010). WRPA paraphrase corpus covers 16 different relations in English and Spanish (Ganitkevitch and Callison-Burch, 2014). PPDB corpus contains over 220 million paraphrase pairs consisting of 73 million phrasal and 8 million lexical paraphrases extracted from 100 million sentence pairs and 2 billion English words by computing distributional

similarity scores using the Google n-grams and the annotated Gigaword corpus (Ganitkevitch et al., 2013). This corpus extended to 23 different languages (Barancíková et al., 2014) later. Vila (Vila et al., 2013) proposed new annotation infrastructure for paraphrase type annotation consisting of an annotation scheme and inner-annotator agreement measures and proofed adequacy and robustness of this infrastructure by annotating three different corpora. Among Turkic languages, to our best knowledge, there is only one Turkish language paraphrase corpus contains 1270 paraphrastic sentences pair covering literary text, movie subtitle, machine translation parallel corpus and news articles (Demir et al., 2012). For Uyghur language, although there are several achievements in morphological analyzing (Wumaier et al., 2009; 麦热哈巴·艾力 et al., 2012), machine translation (Abiderexiti et al., 2013; Mi et al., 2014; Nimaiti and Izumi, 2014) until now there is no reports on constructing Uyghur paraphrase corpus in the literature.

3. Types of Uyghur Paraphrases

In collection of Uyghur paraphrases the first and foremost problem is to define type of Uyghur paraphrase. Uyghur language, like Turkish and Finn belongs to agglutinative language, have vast variety of suffixes. Different suffix connect to the same words construct different meanings of the words, phrases and even sentences. So we referenced the experience of construction of Turkish paraphrase corpus (Demir et al., 2012). Initially we divide Uyghur paraphrase following four categories considering structure, syntactic and computational aspects.

1) Lexical Paraphrases

This type of paraphrases refers to single word paraphrases in the other words, individual lexical items having the same meaning, or synonyms.

For example:

<يۇرت == مەھەللە>		<كۆلەڭگە == سايە>	
Country	Community	Darkness	Shadow

From the example we see that lexical paraphrases are not restricted to strict synonymy. Although يۇرت == مەھەللە country and community are still paraphrases, where country is more general and community is more specific. This paraphrasitc relation can convey approximately the same meaning using different words.

2) Phrasal Paraphrase

This type of paraphrases refers to phrasal fragments sharing the same semantic content. More specifically, it refers multi-word paraphrases, including cases where a single word or single word + inflectional group (IG) maps onto a multiword paraphrase or many-to-many paraphrases.

Example

ياخشى تەربىيە كۆرگەن == ئەخلاقلىق

Be moral Well educated

3) Sentential Paraphrases

A sentential paraphrase refers two sentences that represent the same semantic content.

Example:

نېمىشقا ئۇندىمەيسەن == نېمىشقا زۇۋان سۆزمەيسەن

Why don't you make sound? Why don't you speak?

ئۇ ئاخىر ئاقلىنىپ چىقتى == ئۇنىڭ گۇناھسىزلىقى ئىسپاتلاندى

She/he finally became innocent. Her/his innocence was justified.

Above first example shows that simple sentential paraphrases can be generated easily by simply substituting words or phrases in the original sentences with their respective synonyms or phrasal paraphrase. However, come up with paraphrase like second example is a little difficult for human annotator.

4) Term and its Definition

Terms or words and it's definition or it's approximate same meaning expressed by two or more words.

Example:

پەرز ئىجرا قىلمىسا بولمايدىغان

obligation be indebted for a service

We design paraphrase annotation tool according this four categories.

4. Uyghur paraphrase annotation tools

Because Uyghur online resources are limited, so obtain high qualified paraphrase pairs in unsupervised way (Dolan et al., 2004) or semi supervised way (Ganitkevitch et al., 2013) is not practical. So we developed Uyghur paraphrase annotation tool for constructing

paraphrase corpus by manually with the assistance of the annotation tool. When Demir (Demir et al., 2012) constructing corpus the input texts are different translations of a famous novel, Turkish translations of a foreign movie, Turkish reference translations from an English-Turkish parallel corpus and Turkish articles from a news website. For Uyghur, first two kinds of resources are rare. However, in the same article there is always different description of the same meaning. So the process of using paraphrase annotation tool is that, first, open document using annotation tool, considering human memory mechanism, document usually is not very long, about 300-500 words. And mark paraphrases with assistance of this software. In order to avoid repetitive work of the same annotator we added active learning model, so annotator no need to mark phrases that have already marked before. We implemented active learning model in a very simple way. Every annotated paraphrase was saved in the local access database file. When annotator click "automatic annotate", every candidate paraphrases compared with database in heuristic way. Then display different type of paraphrase with different color.

Uyghur writing system is based on Arabic script, write from right to left. So if we use Latin script annotation mark, much of corpus usually do, there is a little mess with displaying and make annotator confuse. Furthermore, they need to remember what English word or its abbreviation corresponding to Uyghur, those add to additional working time to annotators. In order to avoid these disadvantages we used Uyghur words as annotation mark. The interface of the tools is shown as Fig. 1.

Fig. 1. Interface of Uyghur paraphrase annotation tools

When we construct corpus, we need to calculate work amounts of annotators and annotation agreement so we designed user management components to this tool. This component is responsible for record who annotate which paraphrases and annotated how many.

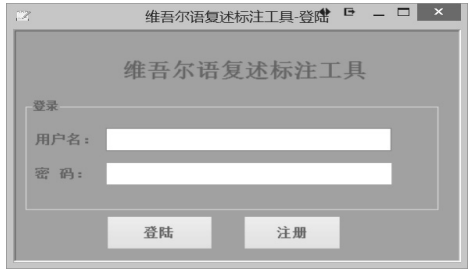


Fig. 2. Login interface

5. Process of Annotation

We ask three undergraduate students from computer science, who taught from elementary school to university in Uyghur language, annotate the same articles. Our intention is to retrieve paraphrase pairs. Not annotate all paraphrases in every article. This means annotator asked do not annotate until he or she fully confident. This elevates work tension and improves annotation agreement.

The annotation process was performed in three ways. The first way includes three steps. The first step is annotators' training. In this step, two annotators independently following examples of paraphrases using our paraphrase annotation tool annotated 10 short articles. Then, problems of the software are solved and disagreements of paraphrases in this 10 articles annotated by this two annotators were discussed with third annotator. The second step is the inner-annotator agreement step. In this step, 150 short articles were selected from Uyghur websites randomly and annotated by two annotators respectively in the assistance of annotation tool. The results were compared with each other and the inter-annotator agreement computed. Regarding the inter-annotator agreement calculation, Kappa measures were used. The kappa score is 0.5 that is below than the desirable score 0.8. So we preceded final step. In the final step, the different annotated paraphrases in second step revised by third annotator to improve the quality of paraphrase

corpus. In this way the overall annotated articles were 201 in which the lexical paraphrases were 572 pairs, phrasal paraphrases were 244 pairs, sentential paraphrase were 114 pairs, terms and its definition is 55 pairs. Although for lexical and phrasal paraphrases statistics we used "pairs", here it means set of two or more than two paraphrases. Because in the overall articles most of the words and phrases were paraphrased more than two equal meaning.

In the second way, we ask the annotators to select sentences from article or book randomly and paraphrase each sentences so that it remains the approximately same information. The students were restricted to only make use of the information and knowledge contained in the sentences and not depend on their commonsense knowledge to alter original meaning of the sentences. In this way we retrieved 2400 sentences. Then in order to research other aspect of paraphrase, 2400 sentences were selected from books, online source and asked annotators to rewrite each sentence but won't consist paraphrase with original sentences. In this way we get balanced positive and negative paraphrases examples for binary classifier for future research.

In the third way, we add translation references to our corpus. In CWMT 2011[1], CWMT 2013 [2]Uyghur -Chinese parallel corpora used for developing Uyghur-Chinese machine translation systems. In these series contests 700 sentences multiple reference translations for CWMT 2011 and 1000 sentences multiple reference translations for CWMT 2013. So we added these 1700 sentences pairs, which consist of a collection of news-related sentences.

6. Conclusion

In this paper we describe the process of constructing Uyghur initial paraphrases corpus. First we define four type of paraphrases, then we annotate or retrieve paraphrase in three ways. In the first way, short papers were chosen from the online source and then annotate paraphrases using Uyghur paraphrase annotation tools. In the second way, in the future in order to build a classifier to predict whether two Uyghur sentences are paraphrases or not, we ask annotator to choose sentences from books or online source and then to paraphrase half of

[1] http://nlp.ict.ac.cn/evalshow.php?id=2011
[2] http://www.liip.cn/CWMT2013/

sentences in correct way, other half of sentences are paraphrased by in a wrong way so that they were consisted similar sentences pair but were not sentential paraphrase each other. Each paraphrases annotated or paraphrased by native speaker who taught Uyghur language from elementary school to University. In the third way, we add 4 different Uyghur translations of the same Chinese sentences to our paraphrase corpus.

7. Future works

In the future we need to expand our corpus improve quality, so that to use machine translation and use the positive, negative paraphrase examples to build automatic classifier that will enable us to retrieve paraphrases from online source automatically.

Acknowledgements

This work has been supported as part of the NSFC (61462083, 60963018, and 61463048), 973 Program (2014cb340506), National Social Sciences Foundation of China (10AYY006), and open project of Xinjiang laboratory of multi-language information technology (049807).

REFERENCES

[1] M. Litip and M. Ablimit, "Method of the management of Uyghur Verbs in Chinese-Uyghur Machine Translation," Journal Of Xinjiang University(Natural Science Edition), vol. 21, no. 1, pp. 77–80, 2004. (in Chinese).

[2] C. Tantug, K. Oflazer and I. D. El-Kahlout, "BLEU+: A Tool for Fine-Grained BLEU Computation," in Proceedings of the Sixth International Language Resources and Evaluation (LREC0'8),Marrakech, Morocco, 2008, pp. 1493–1499.

Abiderexiti, K., Yao, T., Yibulayin, T., Wumaier, A., Yiming, Y., 2013. Implementation of Chinese-Uyghur Bilateral EBMT System, 2013 International Conference on Asian Language Processing (IALP). 2013 International Conference on Asian Language Processing, China,Urumqi, pp. 87–90.

Androutsopoulos, I., Malakasiotis, P., 2010. A survey of paraphrasing and textual entailment methods. J. Artif. Intell. Res. 38, 135–187.

Barancíková, P., Rosa, R., Tamchyna, A., 2014. Improving Evaluation of English-Czech MT through Paraphrasing, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland.

Bhagat, R., Hovy, E., 2013. What Is a Paraphrase? Computational Linguistics 39, 463–472.

Cohn, T., Callison-Burch, C., Lapata, M., 2008. Constructing corpora for the development and evaluation of paraphrase systems. Computational Linguistics 34, 597–614.

Demir, S., El-Kahlout, I.D., Unal, E., Kaya, H., 2012. Turkish Paraphrase Corpus, LREC, pp. 4087–4091.

Dolan, B., Quirk, C., Brockett, C., 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources, Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, p. 350.

Ganitkevitch, J., Callison-Burch, C., 2014. The Multilingual Paraphrase Database, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland.

Ganitkevitch, J., Van Durme, B., Callison-Burch, C., 2013. PPDB: The Paraphrase Database, HLT-NAACL, pp. 758–764.

Madnani, N., Dorr, B.J., 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. Comput. Linguist. 36, 341–387.

Mi, C., Yang, Y., Zhou, X., Wang, L., Li, X., Tursun, E., 2014. A Phrase Table Filtering Model Based on Binary Classification for Uyghur-Chinese Machine Translation. Journal of Computers 9, 2780–2786.

Nimaiti, M., Izumi, Y., 2014. An experiment on Japanese-Uighur statistical machine translation with increased corpus, Cognitive Informatics & Cognitive Computing (ICCI* CC), 2014 IEEE 13th International Conference on. IEEE, pp. 490–495.

Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P., 2010. An evaluation framework for plagiarism detection, Proceedings of the 23rd international conference on computational linguistics: Posters. Association for Computational Linguistics, pp. 997–1005.

Vila, M., Martí, M.A., Rodríguez, H., 2014. Is this a paraphrase? what kind? paraphrase boundaries and typology. Open Journal of Modern Linguistics 2014.

Vila, M., Rodríguez, H., Martí, M., 2013. Relational paraphrase acquisition from Wikipedia: The WRPA method and corpus. Natural Language Engineering FirstView, 1–35.

Wumaier, A., Tursun, P., Kadeer, Z., Yibulayin, T., 2009. Uyghur Noun Suffix Finite State Machine for Stemming, 2009 2nd IEEE International Conference on Computer Science and Information Technology. 2009 2nd IEEE International Conference on Computer Science and Information Technology, Beijing, pp. 161–164.

麦热哈巴·艾力, 姜文斌, 王志洋, 吐尔根·依布拉音, 刘群, 2012. 维吾尔语词法分析的有向图模型. JOURNAL OF SOFTWARE 23, 3115-3129.

TURKIC LANGUAGE SUPPORT IN SKETCH ENGINE

Vít Baisa[a,b], Vít Suchomel[a,b]

[a] NLP Centre, Masaryk University, Brno, Czech Republic
[b] Lexical Computing Ltd, Brighton, UK

Sketch Engine is a corpus manager tool which allows building own text corpora from user-uploaded files or from Internet by downloading and cleaning web pages in a particular language and domain. It also provides many functions to explore the corpus data. We present the level of current support of Turkic languages (namely Azeri, Kazakh, Kyrgyz, Tatar, Turkish, Turkmen, Urdu and Uzbek) in Sketch Engine. It is currently possible to use features of Sketch Engine like concordancing, filtering, sampling, sorting of query searches, wordlist generating, collocation lists extraction, keyword extraction, finding good dictionary examples for words and phrases and some other features.

Additionally, we discuss possible developments for improving Turkic language support in Sketch Engine, starting with incorporating existing tagging tools for Turkic languages, adding terminology extraction and building word sketches and thesauri.We invite Turkic language specialists to join us in our efforts of building large scale and at the same time high quality resources for Turkic languages.

1. Introduction

Turkic language family contains more than thirty languages and the biggest language, Turkish, is spoken by almost 1% of the world population. It is spoken more than Italian or Dutch. Turkish Wikipedia is 10th (Azeri being 39th) biggest measured by the number of active editors[1] however Turkic languages in general are under-resourced from the point of view of corpus linguistics. That is why we have put some effort to creating Turkic language resources and adding at least basic Turkic language support to Sketch Engine. This paper describes the result.

First we describe how we have built several Turkic corpora from Internet. Then we describe current features of Sketch Engine available for these corpora. A discussion of possible other usages follows. At the end we propose possible improvements and future work towards full support of Turkic languages in Sketch Engine.

[1] http://wikistats.wmflabs.org/display.php?t=wp&s=ausers_desc

2. Building Turkic corpora

We selected Turkish, Azerbaijani, Uzbek, Kazakh, Turkmen and Kyrgyz for our 2012 Turkic data collection (Baisa, Suchomel, 2012). The procedure for building general corpora from the web remains the same:

- Start with a small corpus in the target language or create one from Wikipedia texts to build language and encoding detection models.
- Find at least 100 web pages in the target language and use them as starting points for a web crawler.
- Run the web crawler – we have been successfully using SpiderLing, a text corpus oriented crawler (Suchomel, Pomikálek, 2012).
- Alternatively, run WebBootCaT (Baroni et al., 2006), a tool for creating mid-sized corpora from the web using a search engine (the tool is built in the Sketh Engine corpus creation interface) and a part of the Corpus Factory method (Kilgarriff et al., 2010).
- Clean the web data using a set of tools for HTML boilerplate removal, de-duplication (removal of similar sentences, paragraphs or documents) and a robust (web texts aware) tokenizer[1].
- Carry out part-of-speech tagging using a tagger for the target language.
- Store and index the corpus by a corpus manager to allow fast search.

Since a productive inflectional and derivational agglutinative morphology is essential for Turkic languages, any serious corpus based research can benefit from a proper morphological annotation. Although there is not a morphological analyzer built in Sketch Engine, uploading user annotated texts is supported.

Texts in languages written in multiple scripts or spoken in areas of different countries like Tatar and Uyghur are much harder to obtain using the web crawling method. Differences in alphabets and lists of words might be exploited to separate documents in different Turkic

[1] All available for free at http://corpus.tools

languages. Yet, one has to deal with multiple writing systems in the region: Cyrillic, Latin and Arabic.

In case of problems stemming from the issues with crawling, we recommend the search engine driven approach to build large corpora from the web:

- Again, start with a small corpus in the target language or create one from Wikipedia texts.
- Produce a list of words in the corpus sorted by number of occurrences in the corpus from the most frequent word. Use medium frequent words, e.g. from rank 500 to 600 and from rank 1500 to 1600 as seed words for WebBootCaT.
- Let a search engine find web documents in the target language and build the corpus semi-automatically using WebBootCaT.

To gather good quality texts[1] in languages with a scarce Internet presence (which is the case of the most Turkic languages), one can employ less automated means as was shown by (Dovudov et al., 2011):

- Identify Internet sources yielding quality documents, e.g. online newspapers, and government or municipality portals.
- Analyze the web structure of the sources, i.e. locate texts within the site (e.g. find archive of a news site) and determine the important blocks in html pages: this can be automated (Song et al., 2004).
- Write a computer program downloading texts from the web according to findings in the previous step. A recursive run of wget[2] might do the task as well.

Normalization (or unification) of web texts might be required to achieve a good level of quality as reported by (Dovudov et al., 2011):

- Transliteration of letters to the desired script, e.g. from the Latin script or the Arabic script to the Cyrillic script.
- Identification and correction of language specific letters, e.g. replace Н by Ң where appropriate in Kazakh, Kyrgyz, Tatar and Turkmen.

---

[1] A "good quality" text for the purpose of a linguistic research carried on text corpora can be defined as a long sequence of paragraphs of fluent natural sentences.

[2] Wget, a utility for downloading web content, http://www.gnu.org/software/wget/

Table 1

Turkic corpora for language research currently available in Sketch Engine

Language	Name	Corpus size [M tokens]	Lexicon size [M words]	Notes
Azeri	Turkic web – Azerbaijani	115	1.5	Web crawled
Kazakh	Turkic web – Kazakh	175	2.2	Web crawled
Kyrgyz	Turkic web – Kyrgyz	24	0.6	Web crawled
Tatar	Tatar sample	0.29	0.07	Small web corpus gathered using WebBootCaT (Ambati et al., 2012)
Turkmen	Turkic web – Turkmen	3	0.2	Web crawled
Turkish	Turkish WaC	41	1.5	Small web corpus gathered using the Corpus Factory method, parsed with MaltParser[1] (Ambati et al., 2012)
	TrTenTen	4,125	17.2	Web crawled
	OPUS2 Turkish	207	1.5	Parallel corpus[2]
Uzbek	Turkic web – Uzbek	25	0.6	Web crawled

The corpora don't have rich metadata, e.g. domains and text types are missing for all documents. To understand the type of texts in these corpora, it is good to look at the most exploited web domains. In Table 2 you can see top domains for the Turkic corpora.

Table 2

Top domain contained in Turkic corpora

Corpus	Top domains
Turkic web – Azerbaijani	mediaforum.az, az.trend.az, milli.az, mia.az, modern.az, 525.az, ...
TrTenTen	afyonkarahisar.com.tr, savaskarsitlari.org, yeniasya.com.tr, ...

[1] MaltParser, a data driven dependency parser. http://www.maltparser.org

[2] OPUS, the open parallel corpus. http://opus.lingfil.uu.se

Turkic web – Kazakh	alashainasy.kz, egemen.kz, inform.kz, kaz. gazeta.kz, thenews.kz, ...
Turkic web – Kyrgyz	kabar.kg, www.azattyk.org, kg.zpress.kg, erkintoo.kg, ktrk.kg, ...
Turkic web – Turkmen	tmolympiad.org, www.azathabar.org, turkmenistan.gov.tm, cci.gov.tm, ...
Turkic web – Uzbek	uza.uz, shou-biznes.uz, jamiyatgzt.uz, old. uzbekistonovozi.uz, ...

3. Concordances

Query ekmek **2,727** > Shuffle **2,727** (67.27 per million)

First | Previous Page 2 of 137 Go Next | Last

bakterim.net	için bir restorana gittiğinizde masanıza	**ekmek**	istemek zorunda kalabilirsiniz . Biz ekmeği
milligorusportal.com	Bulgur yiyin Bir Fransız kraliçesi vardı .	**Ekmek**	bulamıyorlarsa pasta yesinler demişti .
scribd.com	durum yerlilerin ekim yapmasına engel oldu .	**Ekmek**	olmayınca , Hıristiyanlar , yerlilerin
unknown	basılmış gibi çıkarılmaktadır . BU ülkenin	**ekmeğini**	yiyip , suyunu içip , havasını teneffüs
pdrciyiz.biz	Şu ekin tarlalarını görüyor musun ? Ben	**ekmek**	yemem . Buğday benim hiçbir işime yaramaz
vik2.com	etiketlenmiştir . Etiketler : günü gününe , kuş	**ekmeği**	, pasta , Picrochole , RABELAİS , safra
radikal.com.tr	verilmeye başlandı . Efendim şekerden tut da ,	**ekmeğe**	, benzine , hatta ipliğe kadar karneye
lynchforum.net	aslında , kocasının hamurunu yoğurmaya ,	**ekmeğini**	pişirmeye , evini temizlemeye ve bu gibi
cellotin.com	sahip bir ilişki olduğunu belirtmişlerdir	**ekmek**	kalitesi üzerinde gluteni oluşturan basit
kisiklimahallesi.blogspot.com	sahip oldular . Kısıklı Mahallesi'ne Halk	**Ekmek**	satış noktası açıldı . Ferah Caddesi üzerinde
saglikbilgilerim.com	dü şer . Bu yüzden , kabızlıkta esmer	**ekmek**	yemek daha uygundur . Diyete ; lahana ,
fikiratolyesi.com	bilmediğin konularda konuşmak kabalıktır Mesela	**ekmek**	Kaç çeşit ekmek var Ekmeği nasıl pişirirsin
guneyhaberci.com	yedim . Az veya çok çocuklarıma , evime	**ekmek**	götürdüm . Ama hizmet de verdim . Yani
forum.kanka.net	asla ! Ben kullandıgım benzine , yediğim	**ekmeğin**	fiyatına bakarım , Ben halktan bir vatandaşsam
emlakkulisi.com	kastamonu turhal yozgat çorum çarşamba rüstem	**ekmekçi**	isa gök ak-can akcan ak can pancar kooperatifleri
unknown	yapılmış bir tas ayran veya bir baş soğan	**ekmek**	salmasın hazmedilmesine yardımcı olurdu
agmerkezi.com	çubukları hala yaygın olarak kullanılmaktadır .	**EKMEK**	KIZARTMA MAKİNESİ : 1909 ? da General Electric
ihvanforum.org	dahi bu zorunluluğa tabi , ancak örneğin 5	**ekmek**	fırını tek bir gıda mühendisi istihdam
ozgurokul.org	hastalıklara karşı daha az dirençli olan	**ekmeklik**	buğday , modern piyasa ekonomisine daha
inciturk.com	semboller vaftizde su , Rab bin Sofrasında ise	**ekmek**	ve şaraptır . Sembollerin amacı , sembolize

Fig. 1. Sampled concordance for lemma "ekmek" in TurkishWaC

The main feature available for all corpora is *concordance search*: a powerful full-text search. As many of our Turkic corpora have only word forms (lemmas and other tags are not available), the searching is limited to regular expressions over these word forms. But even with this limitation, the query language (CQL, Corpus Query Language[1]) is expressive enough to allow complex searches.

Once a result is shown, it can be sorted, further filtered (by other CQL queries), randomly sampled (see Figure 1), stored and various frequencies (Figure 2) and visualizations (Figure 3) can be obtained. All these actions can be combined to narrow and fine-tune the original result.

If there are enough hits (examples) in a concordance search, one can extract the most salient collocations from it. The algorithm in Sketch

[1] http://www.sketchengine.co.uk/documentation/wiki/SkE/CorpusQuerying

	word	Frequency
P | N	ekmek	1,406
P | N	ekmeği	260
P | N	Ekmek	232
P | N	ekmeğin	111
P | N	ekmeğini	102
P | N	ekmeğine	71
P | N	ekmeğe	64
P | N	ekmekleri	44
P | N	ekmeklik	37
P | N	Ekmeği	37
P | N	ekmekler	34
P | N	EKMEĞİ	22
P | N	Ekmeğin	21
P | N	EKMEK	20
P | N	ekmeğinin	18
P | N	ekmeklerin	17
P | N	ekmeğimi	15

Fig. 2. The most frequent wordforms of “ekmek” in TurkishWaC

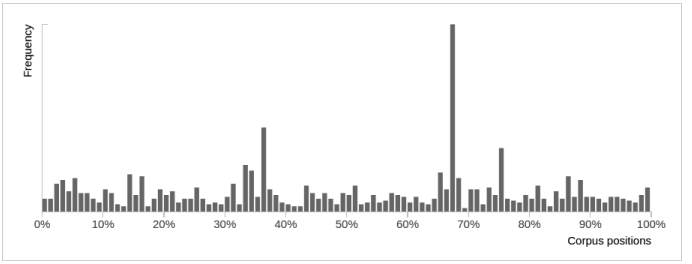


Fig. 3. Frequency distribution of lemma “ekmek” in corpus parts

Engine looks for the most frequent words which co-occur with the searched query and then applies a co-occurrence statistics. We usually use *logDice* (Rychlý, 2008). In Figure 4a you can see collocates derived from the concordance for “ekmek”.

	Frequency	logDice
P | N	fırın	127	10.083
P | N	buğday	113	9.679
P | N	piş	112	9.524
P | N	hamur	70	9.225
P | N	maya	69	9.202
P | N	kepek	50	9.134
P | N	dilim	72	9.096
P | N	kırıntı	40	8.804
P | N	peynir	46	8.695
P | N	makarna	36	8.658
P | N	nohut	36	8.501
P | N	sofra	37	8.295
P | N	şarap	40	8.291
P | N	yufka	25	8.148
P | N	pirinç	29	8.080
P | N	ye	249	7.831
P | N	yemek	80	7.824
P | N	yağ	80	7.784
P | N	kızar	23	7.783
P | N	arpa	24	7.782
P | N	lezzet	28	7.772
P | N	som	19	7.738
P | N	çorba	22	7.663
P | N	lavaş	17	7.661
P | N	mısır	50	7.659
P | N	tereyağ	18	7.562
P | N	tarif	30	7.550
P | N	lokma	17	7.521

Fig. 4. (a) Collocations for lemma “ekmek”

word (lowercase)	Freq
олар	150,215
балалар	60,489
болар	47,628
шаралар	33,832
жағдайлар	23,796
лар	20,844
тұлғалар	15,500
доллар	14,507
Қатысушылар	13,988
оқушылар	13,391
іс-шаралар	13,028
бағдарламалар	12,552
жобалар	12,362
технологиялар	11,824
бұлар	10,644
тауарлар	9,895
компаниялар	8,814
жолаушылар	8,465
толықтырулар	8,189
алар	8,057
оқиғалар	7,878
жазушылар	7,321

(b) Kazakh wordlists for words ending with “лар”

4. Wordlists

Wordlist is another feature universally available for any corpus. Any positional attribute (word, lemma, part of speech, morphology tag, …) can be explored. It is similar to frequency lists of concordance search but wordlists are more general. E.g. you can get the most frequent words, the most frequent lemmas ending with “лар” etc. You may use several constraints and filter the results with regular expressions. You can obtain either raw frequencies or document frequencies per item. In Figure 4b you can see wordlist for all words ending with “лар” from the Kazakh corpus.

5. Word sketches and thesaurus

Word sketches are the core feature of Sketch Engine (hence the name). Word sketch is a one-page, automatic, corpus-derived summary of a word’s grammatical and collocational behaviour. There are several ways to build word sketches.

5.1. CoNLL

We have developed a script which takes CoNLL-annotated corpus as input and generates word sketch grammars[1]. This was also applied on Turkish (Ambati et al., 2012).

5.2. Universal word sketch grammar

We have also processed Turkish part of OPUS parallel corpus using rudimentary tagging (content words, punctuation, numbers) together

kitap **OPUS2 Turkish freq = 10,801** (52.12 per million)

left_content	26,093	1.10	right_content	19,636	0.90	nextleft_content:	9,941	1.00	nextright_content	9,006	1.10
yazarla	28	5.08	fuarına	68	6.57	okuduğum	30	6.09	okuyorum	96	7.50
okuduğum	30	5.01	fuarı	73	6.54	Hesap	30	5.95	okurum	63	7.25
yazdığı	44	5.00	okumak	93	6.28	basılan	17	5.50	fuarına	59	7.24
Hesap	31	4.99	yazmış	73	6.12	kullanılmış	25	5.34	okuyordum	53	7.13
yazılmış	37	4.84	okumayı	56	6.04	yazarla	12	5.17	fuarı	64	7.12
anlatan	33	4.76	yazmak	64	5.83	Kutsal	43	5.00	okuyor	71	6.99
binden	52	4.75	okumaya	42	5.56	çantasındaki	10	4.96	yazmış	79	6.77
basılan	22	4.66	okudum	66	5.54	okuduğun	11	4.90	yazıyorum	58	6.73
Kutsal	43	4.55	tanıtımları	26	5.41	resitalleri	9	4.88	yazdı	85	6.72
Deliler	19	4.51	özeti	28	5.39	yayınevinden	9	4.87	okumak	87	6.64
kullanılmış	25	4.49	sunumları	27	5.39	kütüphanelerinin	9	4.87	okumayı	52	6.63
Interliber	18	4.45	okuyan	38	5.35	Eleştirmenlere	9	4.85	**okudum**	105	6.59

Fig. 5. Universal word sketch for “kitap” in OPUS corpus

[1] http://www.sketchengine.co.uk/documentation/wiki/SkE/SketchesFromCONLL

with so called universal word sketch grammar with very simple rules like "content word to the left from a headword" and other analogous rules. This processing yielded word sketches which can be built also for other Turkic languages but which are not of very high quality and usability. See Figure 5.

5.3. Word sketch grammar

The last and the most advanced way is to write grammar rules manually. It needs both tagged corpus and a language specialist. This is yet to be done.

6. Keyword extraction

If you build your own domain-specific corpus, you can extract keywords from it. The extraction procedure depends on relative frequencies of words in your corpus and in a reference corpus in the same language. For the purpose of this paper we have built a small Turkish corpus using football seed words (a few terms from *Futbal* article on Turkish Wikipedia). Several pages were automatically downloaded and then the corpus was expanded a little with WebBootCat tool, yielding cca 250,000 tokens from football-related Internet pages in Turkish. In Figure 6 you can see the top part of the resulting list of keyword candidates from the domain corpus.

The green keywords were used in building the corpus with WebBootCat. Sketch Engine shows also links to related Wikipedia

Keywords		Score	F	RefF
endirekt	W	1,240.14	481	2,360
vuruş	W	821.79	1,492	26,219
ihlalin	W	768.98	285	2,074
dokunursa	W	636.66	211	1,420
vuruşu	W	571.73	977	24,438
topun	W	512.45	975	27,677
yarda	W	507.78	159	1,116
sportmenlik	W	495.99	164	1,409
atışı	W	487.39	694	19,678
ifab	W	468.18	121	203
kalecinin	W	402.63	321	9,208
vuruşlar	W	386.22	176	3,501
hakemin	W	358.52	516	19,937
yd	W	353.61	148	2,881
ekleminden	W	350.66	90	176
oyuncuya	W	350.20	570	23,086
ihlalden	W	341.94	103	921

Fig. 6. Keyword extraction from a domain-specific (football) corpus

articles (Turkish Wikipedia in this case). The score expresses how salient a keyword is in the domain corpus when compared with a general (much bigger) Turkish reference corpus. The last two columns are raw frequencies in the focus and in the reference corpus. It is also important to note that neither of the authors has any knowledge of Turkish language thus it is possible that the keywords are not perfect. The same methods could be used to build e.g. Tatar corpus and extract keywords from it as it is fully statistically-based approach. More info about the extraction procedure can be found in (Kilgarriff, 2014).

7. Further work and development

The support for Turkic language can be substantially improved. The two most beneficial improvements are discussed below.

7.1. Term extraction

Recently we have developed term extraction for several languages: English, Spanish, German, Czech and a few others (Kilgarriff, 2014). To add a new language to the list, it is necessary to describe possible terms (usually noun phrases) using advanced CQL queries. These queries both describe the grammar rules for matching all possigle term phrases but also they describe how the resulting basic word form for terms should look like.

7.2. Morphological analyzer integration

Advanced Sketch Engine features, such as word sketches and thesaurus, or querying the corpus for morphological categories require a morphologically annotated corpus. Although annotated texts can be loaded into the Sketch Engine, it would be much more convenient for anyone building a Turkic corpus if the tool made the tagging for them.

The requirements for embedding a morphological analyzer in the corpus building interface are:

- Software running in a Unix-like environment.
- Command line interface for batch processing of large quantities of data.
- Documentation: evaluation of the tagger, description of possible output tags.
- Licence allowing to incorporate the tool in Sketch Engine.

8. Conclusion

We have described the current support of Turkic languages in Sketch Engine. It enables a basic analysis and users can upload preprocessed data and use all the standard features of Sketch Engine. With this paper we hope to attract Turkic language specialists to use this powerful tool for exploring the richness of all Turkic languages. Sketch Engine is currently used at many language institutions in Europe and we think that it can boost language research of Turkic languages, its lexicography, terminology and linguistics in general.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013.

REFERENCES

Baisa, V., Suchomel, V. (2012). *Large corpora for Turkic languages and unsupervised morphological analysis.* In *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12), Istanbul, Turkey. European Language Resources Association (ELRA)*.

Suchomel, V., Pomikálek, J. (2012). *Efficient web crawling for large text corpora*. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*. pp. 39–43.

Baroni, M., Kilgarriff, A., Pomikálek, J., Rychlý, P. (2006). *WebBootCaT: a web tool for instant corpora.* In *Proceeding of the EuraLex Conference*, pp. 123–132.

Dovudov, G., Pomikálek, J., Suchomel, V., Šmerk, P. (2011). *Building a 50M Corpus of Tajik Language.* In *RASLAN 2011 Recent Advances in Slavonic Natural Language Processing.*

Rychlý, P. (2008). *A lexicographer-friendly association score.* In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pp. 6–9.

Song, R., Haifeng L., Ji-Rong W., Wei-Ying M. (2004). *Learning block importance models for web pages.* In *Proceedings of the 13th international conference on World Wide Web*, pp. 203–211. ACM.

Kilgarriff, A., Reddy, S., Pomikálek, J., Avinesh, P. V. S. (2010). *A Corpus Factory for Many Languages.* In *LREC 2010*.

Kilgarriff, A., Jakubíček, M., Kovář, V., Rychlý, P., Suchomel, V. (2014). *Finding terms in corpora for many languages with the Sketch Engine*. EACL 2014, 53.

Ambati, B. R., Reddy, S., Kilgarriff, A. (2012). *Word Sketches for Turkish*. In *LREC.* pp. 2945–2950.

Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D. (2004). *The Sketch Engine*. Information Technology. 2004.

MULTILINGUAL DATABASE OF TURKIC COLOR NAMES: STRUCTURE AND DESIGN

Ayrat Gatiatullin, Marat Kurmanbakiev, Bulat Khakimov[1]

Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan Federal University
Kazan, Tatarstan, Russia
[1]khakeem@yandex.ru

Multilingual electronic databases of certain vocabulary domains, such as color names (coloratives), represent a valuable resource for theoretical and applied studies in various fields. We can mention lexicology, lexicography, general linguistics, orthography, ethnopsychology, machine translation and information retrieval, educational technology, as well as creating automated workspaces for various specialists (in this case, for example, painters, cartoonists, designers, etc.). Linguistic database being developed is intended for presentation, storage and efficient use of the fullest possible information about the structure and semantics of coloratives in Tatar and other Turkic languages. The structural-conceptual model of the database is based on studies in the field of cognitive linguistics, dialectology, phraseology and folklore on the material of dictionaries and corpus data. At this stage of the project, 271 different colorative was determined for Tatar language. The structure of the database reflects derivational and grammatical features of Turkic coloratives, particularly their derivational capacity and inflection, so it describes adjective, noun and verb derivatives of lexems with the meaning of color. The database model also includes information about dialectal variants of coloratives, idioms and folklore texts containing coloratives, as well as their cognitive and denotative relationships. The database is populated through special developed software that can be accessed via Internet. The database structure enables us to include new languages and features easily. At the same time it allows to simplify the population process and make search more convinient. Further development of the project involves population of the database for other Turkic languages, and improvement of the search functionality. The authors hope that their project could be a useful resource for comparative research and applied developments in the field of computer-based processing of Turkic languages.

1. Введение

Электронные многоязычные базы данных таких отдельных классов лексики, как цветообозначения (колоративы), представляют собой ценный ресурс для теоретических исследований и

прикладных разработок в различных областях. В первую очередь, это лексикология, лексикография, общее языкознание, орфография, этнопсихология, машинный перевод и информационный поиск, образовательные технологии, системы автоматизации рабочих мест специалистов (в данном случае, например, художника, мультипликатора и др.).

Разрабатываемая лингвистическая база данных тюркских цветообозначений предназначена для представления, хранения и эффективного использования как можно более полной информации о структуре и семантике колоративов в татарском и в других тюркских языках.

Разработка структурно-концептуальной модели базы данных была основана на исследованиях в области когнитивной лингвистики, на материале толковых словарей, корпусных данных, а также исследований по диалектологии, фразеологии, фольклористике и др. В общей сложности, на данном этапе для татарского языка был выделен 271 колоратив различной структуры.

В структурно-концептуальной модели лингвистической базы данных выделяются две основные составляющие: концептуальная модель и структурная модель. Концептуальная модель включает в себя все концептуальное пространство понятия (все его лексические значения, значения семантического гнезда, когнитивные и денотативные связи), а структурная рассматривает внешнее строение лексем и словообразовательные механизмы.

2. Структурный компонент модели базы данных

В плане структуры базы данных цветообозначения анализируются с традиционной точки зрения, а именно – их деривационного потенциала, то есть анализируются адъективные, субстантивные и глагольные дериваты слов, называющих цвета.

Так, суффиксальный способ словообразования адъективных колоративов характеризуется структурой: *основа*+*-лЫ*. Модель *прил.*+*-лЫ* выражает значение ‘имеющий тот цвет, который обозначен производящей основой’ (условно называется в базе данных как “обладательное прилагательное”).

Кроме собственно словообразовательных особенностей, в базе данных для прилагательных-цветообозначений отражены

грамматические формы, выражающие степень и интенсивность признака (усиление или ослабление), а также парные сочетания адъективов на –лЫ.

В таблице 1 представлена структура и примеры заполнения таблицы базы данных “Признаки”, содержащей дериваты и грамматические формы прилагательных-цветообозначений.

Таблица 1

Признаки

Наименование параметра	Примеры	Перевод
1.1. Обладательное прилагательное (ОП)	аклы	‘с белым’
1.1.1. Сочетание ОП	аклы-каралы	‘черно-белый’
1.2. Усиление цвета		
1.2.1. Сравнительная степень	аграк	‘белее’
1.2.2. Превосходная степень	ап-ак	‘белый-пребелый’
1.3. Ослабление		
1.3.1. -сУ	зәңгәрсу	‘голубоватый’
1.3.2. -сЫл	аксыл	‘беловатый’
1.3.3. -ГЫлт	кызгылт	‘красноватый’
1.3.4. -ЫлҗЫм	күгелжем	‘синеватый’
1.3.5. –бҮз	акбүз (тат.) Ақбоз (каз.)	‘бело-сивый’

Глагольные дериваты, образованные от основ со значением цвета, в тюркских языках сравнительно немногочисленны. Как правило, это связано с тем, что единственное значение, которое они могут выражать – это проявление цвета основы и различные связанные метафорические значения.

Структура и примеры заполнения соответствующей таблицы базы данных “Действия” представлены в таблице 2.

Таблица 2

Действия

Наименование параметра	Примеры	Перевод
2.1. -лА	акла	
2.1.1. Залоговые формы		
2.1.1.1. -Ыл	ақтал (каз.)	

2.1.1.2. -Ын	аклан aklan (тур.) актан (кир.)	‘оправдываться’, ‘реабилитироваться’, ‘подтверждаться’
2.1.1.3. -Ыш	аклаш	‘помолвка’ (устар.)
2.1.1.4-ДЫр	аклат	
2.2. -Ар	агар	‘белеть’
2.2.1. Залоговые формы		
2.2.1.1. -Ыл		
2.2.2.2. -Ын	агарын	‘побелеть’, ‘побледнеть’ (о цвете лица)
2.2.2.3. -Ыш		
2.2.3.4. -ДЫр	агарт	‘белить’, ‘просвещать’ (перен.)

Большая часть субстантивных дериватов цветообозначений, представленная в толковых словарях, образована путем сложения двух основ, одна из которых обозначает предмет или явление, которое характеризует колоратив. В этих случаях лексемы четко разделяются на две составные части, имеющие самостоятельные значения. Например, в современном татарском языке имеется большой пласт сложных существительных, возникших от атрибутивных словосочетаний. В основном, в них используются колоративы: *ак сакал* ‘белая борода’ − *аксакал* ‘аксакал’, *ак чәчәк* ‘белый цветок’ – акчәчәк ‘ромашка’, *ак кош* ‘белая птица’ – *аккош* ‘лебедь’, *ак күбәләк* ‘бабочка’ – *аккүбәләк* ‘капустница’. Большинство таких субстантивных дериватов являются названиями объектов живой природы – растений и животных.

Субстантивных дериватов, образованных суффиксальным способом, относительно немного. В основном, они представлены моделью *основа+-лЫк*, которая выражает название цвета, обозначенное производящей основой. Эта форма есть практически у всех колоративов в татарском языке: *кызыллык* ‘краснота’, *аклык* ‘белизна’, *сарылык* ‘желтизна’.

Структура и примеры заполнения таблицы базы данных “Именные конструкции” представлены в таблице 3.

Таблица 3

Именные конструкции

Наименование параметра	**Примеры**	**Перевод**
3.1. Образованные с помощью аффиксов		
3.1.1. -лЫк	аклык, кызыллык	‘белизна’, ‘краснота’
3.1.2. -чА	яшелчә	‘овощи’, ‘зелень’
3.2. Образованные с помощью корневых морфем		
3.2.1. Частотные в разных языках		
3.2.1.1. –баш (‘голова’)	акбаш (тат.)	‘тысячелистник’
	Ақбас (каз.)	‘мокрец’
	Akbaş (тур.)	‘морской гусь’
	Акбаш (кир.)	‘мокрец’
3.2.1.2. –күз (‘глаз’)	аккүз	‘белоглазка’
3.2.1.3. –тамыр (‘корень’)	актамыр	‘пырей’
3.2.2. Уникальные нарицательные		
	акбәкәл	‘белоногий конь’
	акбур	‘мел’
	аккош	‘лебедь’
	аккургаш	‘олово’
	аклан	‘светлая поляна’, ‘прогалина’
	акчарлак	‘чайка’
3.2.3. Имена собственные	Акбатыр	
	Акбәк	

В отдельную таблицу базы данных выделены составные конструкции. Поскольку для автоматической обработки языковых данных важную роль играют формальные признаки, в эту таблицу включаются как именные, так и глагольные конструкции, которые выражают единое значение, но согласно орфографической традиции пишутся раздельно. Иные композитные дериваты со слитным написанием (“сложные”) и написанием через дефис (“парные”) включены в соответствующие разделы базы данных на основе их частеречной принадлежности.

В таблице 4 показаны структура и пример заполнения таблицы базы данных “Составные конструкции”.

Таблица 4

Составные конструкции

Наименование параметра	Примеры	Перевод
Составные имена		
алтын (‘золото’)	ак алтын	‘платина’
балчык (‘глина’)	ак балчык	‘каолин’
Составные глаголы		
алыштыру (‘обмениваться’, ‘менять’)	ак алыштыру	‘обмен подарками при помолвке’
төшү (‘падать’, ‘опускаться’)	(күзгә) ак төшү	‘появление бельма на глазу’

3. Реализация базы данных

Для заполнения базы данных цветообозначений разработано специальное программное обеспечение, доступ к которому осуществляется с помощью веб-сайта. Рассмотрим техническую реализацию этой программы. Данные, с которыми необходимо проводить операции хранятся в таблицах базы данных MySQL на сервере Академии наук Республики Татарстан, которая и составляет основную часть системы. Доступ к данным из БД осуществляется через Интернет с помощью PHP-скриптов. В качестве тонкого клиента для доступа к системе используется веб-браузер.

Преимущества данного типа систем заключаются в том, что система не требует установки специальных компонентов на пользовательскую машину, а также не зависит от конфигурации пользовательской машины. Все операции с данными выполняются на сервере, а значит, при разработке системы необходимо учитывать только его конфигурацию. Общая архитектура системы представлена на рис.1.

Интерфейс системы рассчитан на пользователей двух уровней: конечные пользователи и эксперты. Конечные пользователи получают доступ только к поисковой части системы, в которой они могут осуществлять различные запросы к базе данных. Поисковая часть системы на схеме обозначена значком ‘Поиск’.

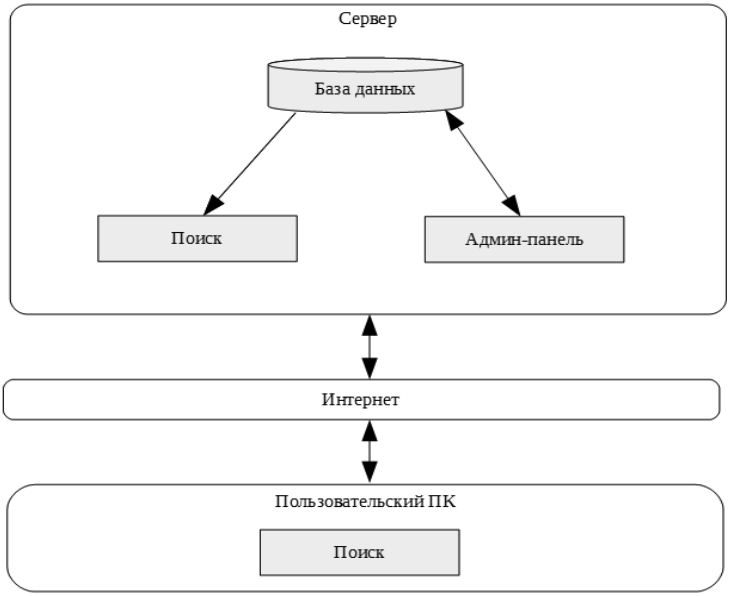


Рис. 1. Архитектура системы

К категории экспертов относятся специалисты, которые могут осуществлять заполнение базы данных. Для них предназначена админ-панель, доступ к которой осуществляется с помощью пароля. Админ-панель позволяет зарегистрированным пользователям, имеющим определенную категорию доступа, вносить изменения в базу данных и в конфигурацию самой системы.

При регистрации эксперты могут получить разный статус: администратор - имеет права на удаление, добавление, редактирование всех записей для одного или нескольких языков, просмотр предлагаемых изменений, сделанных пользователями ранга «Редактор» для своих языков.

• редактор - имеет права на добавление записей в базу данных, а также редактирование записей для одного или нескольких языков. Все изменения, сделанные пользователем данного типа, сохраняются в таблицу с временными записями, и переводятся в основную базу данных системы только после проверки пользователями категорий «Администратор».

Абстрактная архитектура базы данных представлена на рис. 2.

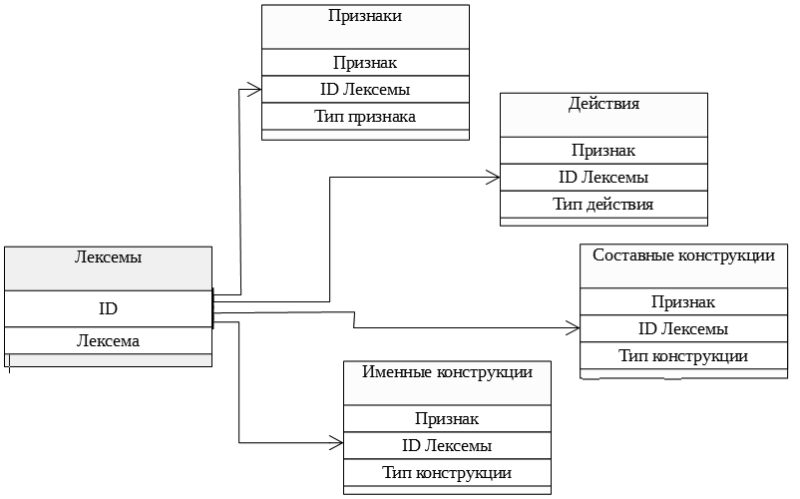


Рис. 2. Архитектура БД

Как показано на рис.2, в БД имеется пять основных таблиц с требуемыми данными. Таблица лексем содержит своеобразный набор ключей, которые определяют принадлежность того или иного свойства к определенной лексеме.

Такая структура позволяет, как облегчить процесс наполнения базы данных, уменьшая вероятность совершения ошибок, и сделать поиск по базе более удобным, так и минимизировать изменения в программной части скриптов и в структуре БД, в случае расширения проекта новыми языками и признаками.

4. Заключение

Модель базы данных тюркских цветообозначений, помимо информации о деривационной структуре колоративов, включает также данные о диалектных вариантах, употреблении в составе фразеологизмов и фольклорных текстов, когнитивных и денотативных связях. Дальнейшее развитие проекта предполагает заполнение базы данных для других тюркских языков, совершенствование механизмов поиска и структуры запросов. Авторы надеются, что разрабатываемая лингвистическая база данных станет полезным

источником для сравнительных исследований и прикладных разработок в сфере компьютерной обработки тюркских языков.

ЛИТЕРАТУРА

Berlin, B., & Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. Berceley-Los Angeles: University of California Press.

Wierzbicka, A. (2008). Why there are no 'colour universals' in language and thought? *Journal of the Royal Anthropological Institute,* Vol. 14, Issue 2, pp. 407–425.

Galieva, A.M., Sitdykova, A.F., & Suleymanov, D.Sh. (2014). Principles of development of the linguistic database of Tatar colour terms.

Кононов, А.Н. (1975). Семантика цветообозначений в тюркских языках. *Тюркологический сборник*. М.: Наука. С. 159–179.

Бакиров, М.Х. (2015). Семантика цвета в контексте истории, искусства и художественной словесности тюрков. *Филология и культура. Philology and Culture*. №1(39). Казань. С.114–119.

Ситдыкова, А.Ф. (2014). К созданию структурно-концептуальной модели для лингвистической базы данных цветообозначений татарского языка. *Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2014*. Казань: "Фэн". С. 230–235.

Сулейманов, Д.Ш., Ситдыкова А.Ф. (2014). Принципы создания лингвистической базы данных по татарским цветообозначениям.

Татар теленең аңлатмалы сүзлеге (2005). Баш ред. Ф.Ә. Ганиев. Казан: «Матбугат йорты».

Татар теленең аңлатмалы сүзлеге (1977-81). Редкол. Л.Т. Мәхмүтова, М.Г. Мөхәммәдиев, К.С. Сабиров, Ш.С. Ханбикова. Казан: Татар. кит. нәшр. 3 томда.

Татарско-русский словарь (2007). Казань: Магариф. В 2-х т.

Татар теленең диалектологик сүзлеге (2009). Төз. Ф.С. Баязитова, Д.Б. Рамазанова, З.Р. Садыйкова, Т.Х. Хәйретдинова. Казан: Татар. кит. нәшр.

Татарско-русский словарь однокоренных слов (2007). Сост. Ф.С. Сафиуллина, З.Н. Кириллова, А.Ш. Юсупова, И.Г. Мифтахова. Казань: Казанский государственный университет.

Татарский национальный корпус "Туган тел": http://web-corpora.net/TatarCorpus/search/?interface_language=ru

THE NEWSPAPER CORPUS OF THE YAKUT LANGUAGE

Nyurgun Leontiev

a North-East Federal University, Kulakovskogo 48,
Yakutsk, 677013, Russia
leonza@mail.ru

This article describes the newspaper corpus of the Yakut language. The linguistic corpora of the Yakut language were developed only relatively recently. The newspaper corpus was created for development of automatic natural language processing. In article is described the using corpus for language identification. For creating the corpus above 21 thousand newspapers' articles were collected. Total number of words of newspaper corpus is about 12 millions words. The process of lemmatization of newspaper corpus was done taking into consideration main features of the Yakut language. The author has developed frequency tables of language with using N-gramms. Several vocabularies of the Yakut language were developed as a result of the analysis of newspaper corpus. This paper shows the result of the last two years work.

1. Газетный корпус

Корпусная лингвистика якутского языка начала развиваться сравнительно недавно. Коллективом под руководством Заморщиковой Л.С. был создан ассоциативный словарь якутского языка (Заморщикова, 2014). Институт гуманитарных исследований и проблем малочисленных народов Севера участвует в работах по разработке параллельного корпуса. Имеются некоторые работы, которые имеют некоторый объем в рамках работы института языкознания РАН.

Автором были созданы требования для создания корпуса на основе машинных корпусов английского и русского языков (Леонтьев, 2014), но оказалось что эти требования недостаточны для корпуса якутского языка. Количество словоупотреблений корпуса были расширены до 12 млн. слов, что дало более объемный результат по количеству словоформ.

Для корпуса были собраны статьи из республиканских газет на якутском языка, таких как «Кыым», «Саха Сирэ» и Интернет-газет «Аартык.ру» и «Sakhalife.ru». Период выпуска газет от 2006 года до 2015 года.

Таблица 1

Количественный состав газетного корпуса

Название	Количество статей	Количество слов
Кыым	8870	5031348
Саха Сирэ	11763	7092159
Аартык.ру	588	309884
Sakhalife.ru	269	149669
Всего	21490	12583060

Статьи собирались в автоматическом режиме, с применением программного определителя языка.

2. Обработка корпуса

В ходе сбора материала выявились проблемы исходящие из-за использования шрифтов разных разработчиков, которые в отсутствие стандарта Unicode внедряли собственные шрифты с разнобойными кодировками символов для использования в кодировке Windows-1251. Все это привело к тому, что один и тот же знак мог иметь разную позицию в шрифтах, или же был исправлен символ, принадлежащий другой кодировке.

В ходе обработки корпуса был собран словарь словоформ – 380 тыс. слов. Для тюркских языков такое количество словоформ будет недостаточным, так как от одной леммы можно образовать несколько тысяч словоформ.

Были выявлены частотные зависимости для слов, символов якутского языка, в том числе и с учетом диграфов и дифтонгов. Составлены частотные словари: символьные, словоформ и N-грамм (Протопопова, 2014). Также составляется словарь имен собственных с разделением на фамилии, имена и географические названия.

Отдельной строкой идет идентификация языка текста в автоматическом режиме. Были испробованы словарный способ, способ с помощью большого словаря словоформ и с помощью N-грамм (биграмм и триграмм). Наилучший результат по быстроте и более-менее нормальный результат по качеству показал метод триграмм.

Для автоматизированной лемматизации был разработан программный модуль с применением базы аффиксов и правил, что

позволило автоматизировать процесс лемматизации, хотя и встречаются неоднозначности лемматизации, которые необходимо решить (Леонтьев, 2015). В ходе работы была создана база данных аффиксов якутского языка по результатам обработки газетного корпуса, но база не охватывает все аффиксы якутского языка.

В ходе лемматизации были выявлены варианты фонетизации заимствованных слов, для слов с неустановившийся фонетизацией были выявлены несколько вариантов правописания. Проблема фонетизации заимствованных слов должна быть решена мерами согласования языковых норм, так как правописание таких слов является сложной задачей.

Заключение

В дальнейшей планируется закончить автоматическую и ручную разметку газетного корпуса, а также доработать программные средства для автоматической обработки текста. Необходимо продолжить сбор первичных материалов из сайтов газет, а также из бумажных носителей. Объем работы очень большой и трудоемкий, и поэтому необходимо доработать автоматизированные методы анализа якутского языка.

ЛИТЕРАТУРА

Заморщикова Л.С. *Ассоциативный тезаурус якутского языка* // Гуманитарные научные исследования. – 2014. – № 2 [Электронный ресурс]. URL: http://human.snauka.ru/2014/02/6027 (дата обращения: 19.06.2015).

Леонтьев, Н.А. *Национальный корпус якутского языка – технический подход* // Труды конференции TEL-2014, Казань, – стр. 122–124.

Протопопова В.Ф. *Частотная таблица символов якутского языка с учетом диграфов и дифтонгов*//Информационно-телекоммуникационные системы и технологии (ИТСиТ-2014): Материалы Всероссийской научно-практической конференции. Кузбас. гос. техн. ун-т им. Т.Ф. Горбачева. – Кемерово, 2014. – С. 141–142.

Леонтьев Н.А. *Идентификация языка текстового сообщения с помощью газетного корпуса якутского языка* // Universum: Технические науки: электрон. научн. журн. – 2014. – № 8 (9)./ URL:http://www.7universum.com/en/tech/archive/item/1539 (дата обращения: 10.11.2014).

Леонтьев Н.А. *Вопросы автоматизированной лемматизации якутского языка* // ФӘН-НАУКА. Бугульма, – 2015, №2 (41) – С. 15–16.

SEARCH ENGINE FOR THE "TUGAN TEL" TATAR NATIONAL CORPUS: MAIN DECISIONS

Olga Nevzorova[1], Damir Mukhamedshin, Ruslan Bilalov

Research Institute of Applied Semiotics of the Academy of Sciences of Tatarstan Republic, Kazan, Russia
Kazan (Volga Region) Federal University, Kazan, Russia
[1]onevzoro@gmail.com

This paper presents a report on the research and development of Tatar corpus management system. Our project is based on the famous program solutions. First, it is Eastern Armenian National Corpus (http://eanc.net), which primary provides a platform for the electronic corpus of the Tatar language "Tugan Tel" (http://web-corpora.net/TatarCorpus). It has 5 different search types: direct search by word, direct search by lemma, reverse search, accurate and inaccurate search. Functionality implemented in the platform of East Armenian National Corpus, is the basis for the Tatar corpus manager. Another example is the Russian National Corpus (http://ruscorpora.ru). A distinctive feature of this platform is the support for extended syntax of search queries for direct searching, namely, the support for negative keywords, the search by parts of words and Boolean operators. These features also have been added to the search functionality of the Tatar corpus manager. In addition, the Tatar corpus manager implements the functionality of phrasal search, using arbitrary morphological formulas with operators AND, OR, NOT and parentheses for prioritization and identification of logical errors in formulas.

Corpus data are presented in the form of the semantic structure of classes. Such structure allows to perform searching for many lexical and morphological parameters, as well as finding logical errors in user search queries. Total count of word forms in database is over 46 million.

The system architecture and, in particular, the database have been constructed in such way as to be able to handle the requests of the following types: direct search (by word form or lemma); reverse search (by morphological properties). Reverse search supports the following types of logical formulas: 1) with conjunctions, 2) with disjunctions, 3) with negations, 4) arbitrary logical formulas (using conjunction, disjunction and negation) and 5) mixed search (by morphological properties and word form or lemma).

The main advantages of developed corpus manager are the supporting the Tatar language implementation and the possibility of rapid integration with electronic corpuses of other languages, the identification of logical errors, and the openness of used technologies. Testing the system has

shown that the time required for processing and executing the search query by the system is not more than 0.05 seconds in 98.71% of cases the lexical search, in 77.71% of cases the morphological search, and in 98.08% of cases the lexical-morphological search. In many ways, these results were achieved through the presentation of data in a semantic network and the proposed methods of representation and search query processing.

1. Татарский национальный корпус “Туган тел”

Тюркская корпусная лингвистика в настоящее время достаточно активно развивается. Из известных проектов электронных корпусов тюркских языков можно отметить электронные корпусы турецкого (Aksan, Y. et al, 2012; Dalkiliç, G. and Çebi Y., 2002; Say et al, 2002; METU Turkish Corpus), уйгурского (Yusup Aibaidulla and Kim-Teng Lua, 2002), башкирского (Бускунбаева Л.А., 2011), хакасского (Шеймович, 2011), казахского (http://til.gov.kz), тувинского (Салчак, 2012) языков. Татарский национальный корпус «Туган тел» является лингвистическим ресурсом современного литературного татарского языка. Объем корпуса на июнь 2015 года составляет более 46 миллионов словоформ. Корпус содержит тексты различных жанров (художественная литература, тексты СМИ, тексты официальных документов, учебная литература, научные публикации и др.). Каждый документ имеет метаописание (автор, название, выходные данные, дата создания, жанр и др.). Тексты, включенные в корпус, снабжены морфологической разметкой (информация о части речи и грамматических характеристиках словоформы). Морфологическая разметка текстов корпуса выполняется автоматически с использованием модуля двухуровневого морфологического анализа татарского языка, реализованного в программном инструментарии PC-KIMMO.

2. Прототип корпус-менеджера для Татарского национального корпуса «Туган тел»

В настоящее время разработаны различные по функционалу системы корпус-менеджеров, предназначенных для обработки электронных корпусов текстов. Для Татарского национального

корпуса «Туган тел» (ТНК ТТ) нами разработан корпус-менеджер, реализующий развитый поисковый функционал и семантические модели представления корпусных данных (Невзорова и др. 2015а, 2015б). Наиболее близкими проектами являются корпус-менеджеры Восточно-армянского национального корпуса и Национального корпуса русского языка. Восточно-армянский национальный корпус (http://eanc.net), на платформе которого был первоначально размещен электронный корпус татарского языка «Туган Тел» (http://web-corpora.net/TatarCorpus), имеет 5 различных видов поиска: прямой по иск по слову, прямой поиск по лемме, обратный поиск, точный и неточный поиск. Функционал, реализованный в платформе Восточно-армянского национального корпуса, является базовым для корпус-менеджера, представленного в настоящей статье.

Национальный корпус русского языка (http://ruscorpora.ru) обладает функционалом, схожим с платформой Восточно-армянского национального корпуса. Отличительной особенностью платформы Национального корпуса русского языка является поддержка расширенного синтаксиса поисковых запросов при прямом поиске, а именно поддержка минус-слов, поиска по части слова и логических операторов. Эти возможности были добавлены в поисковый функционал системы ТНК ТТ. Кроме того, в системе ТНК ТТ реализован фразовый поиск.

Помимо разработки расширенного поискового функционала перед авторами стояли задачи оптимизации времени исполнения поисковых запросов (менее 1 секунды на запрос), поддержка произвольных морфологических формул с использованием операторов И, ИЛИ, НЕ и выставления приоритетов выполнения при помощи скобок, а также выявление логических ошибок в формулах. Примером логической ошибки является противоречивая формула «!(N|V),INF_1», которая означает «НЕ имя существительное (N) И НЕ глагол (V) И инфинитив, оканчивающийся на аффикс *–ырга* (INF_1)». Противоречивость данной формулы заключается в том, что все элементы, относящиеся к классу «INF_1» также относятся и к классу «V», но в первой части все элементы класса «V» исключаются, соответственно, результатов по данному поисковому запросу существовать не может.

2. Архитектура корпус-менеджера

Для функционирования корпус-менеджера используется следующее открытое программное обеспечение: веб-сервер Apache (или HHVM), интерпретатор PHP, СУБД MariaDB, in-memory хранилище Redis (кэширующий сервис), сервер очереди MemcacheQ. Всё программное обеспечение распространяется с открытым исходным кодом и может быть свободно использовано в некоммерческих целях.

Программная реализация использует концепцию MVC (Model-View-Controller), несколько измененную для решения задач корпус-менеджера.

Рис. 1. Архитектура системы.

На рисунке 1 представлена общая концепция архитектуры системы на примере поискового функционала. После формирования запроса пользователем, первым этапом является проверка запроса подходящим контроллером на наличие синтаксических ошибок, а также подозрительных действий со стороны пользователя (атаки). Вторым этапом является обработка поискового запроса, а именно его проверка на наличие логических ошибок и приведение текстового запроса к структурированному объекту запроса. На третьем этапе объект запроса используется моделью поиска для формирования запроса к БД. Затем на основе сформированного запроса к БД проверяется наличие кэшированных результатов, при их наличии дальнейший запрос к БД не производится, что существенно

уменьшает время на выполнение запроса и потребляемые ресурсы. На пятом этапе выполняется запрос к БД, а полученные по этому запросу результаты используются для формирования объектов поисковой выдачи. Последним этапом является отображение результатов поиска пользователю. Объекты поисковой выдачи выводятся в браузер пользователя в виде документа HTML.

3. Представление данных в системе корпус-менеджер

В системе корпус-менеджер данные представляются в виде семантической структуры классов, которая представлена на рисунке 2. Базовым классом является класс Документы, включающим в себя подкласс Контексты, который в свою очередь включает в себя подкласс Разборы. Последний подкласс содержит в себе около 55,7 млн. элементов, делится на несколько подклассов по морфологическим признакам, а также связан с классами Словоформы, Леммы, Морфологические признаки. Между элементами классов Разборы, Контексты и Документы имеются связи «часть-целое» (на рисунке сплошная линия); между элементами класса Словоформы (и Леммы) и Разборы – связи «часть-целое»; между элементами классов Разборы и Морфологические признаки – мно-

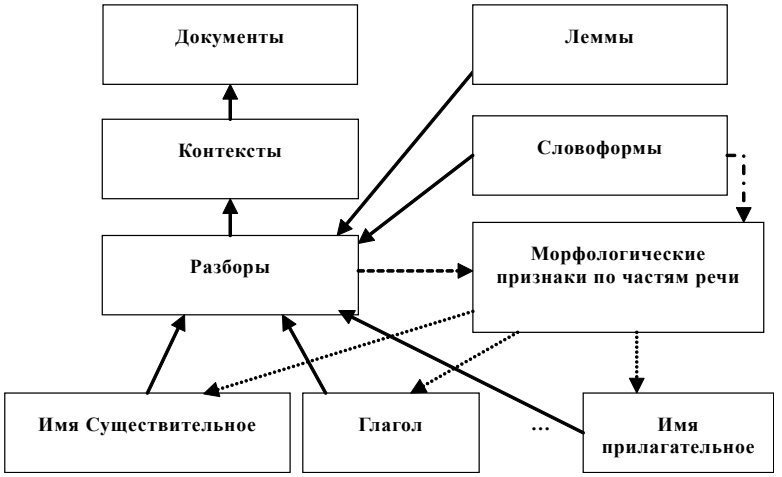


Рис. 2. Связи между элементами в системе.

жественные связи «hasGrammaticalFeature» («имеет морфологический признак», на рисунке квадратные точки); между элементами классов Словоформы и Морфологические признаки – множественные связи «usedWithGrammaticalFeature» («употребляется с морфологическим признаком», на рисунке штрих-пунктир); между элементами класса Морфологические признаки и подклассами класса Разборы – связи «isFoundIn» («встречается в», на рисунке круглые точки).

Такая структура позволяет производить поиск с учётом многих лексических и морфологических параметров, а также находить логические ошибки в пользовательских поисковых запросах. Каждый элемент, кроме представленных связей имеет определенные свойства, такие как название, позиция в контексте, позиция в документе и т.п. Эти свойства не используются непосредственно в процессе поиска, но могут быть выведены пользователю наряду с остальными.

4. Хранение данных электронного корпуса

Архитектура системы и, в частности, БД построена таким образом, чтобы иметь возможность обрабатывать запросы следующих типов [1]:

- Прямой поиск (по словоформе или лемме);
- Обратный поиск (по морфологическим свойствам), морфологические свойства могут быть представлены в виде формулы одного из видов:
  - С использованием конъюнкций;
  - С использованием дизъюнкций;
  - С использованием отрицаний;
  - Произвольная формула (с конъюнкцией, дизъюнкцией и отрицанием);
  - Смешанный поиск (по морфологическим свойствам с указанием словоформы или леммы).

Размеченные тексты разбираются на отдельные словоформы и их разметку, затем записываются в БД в виде обратного индекса, списка предложений (контекстов) и списка документов (рис. 3). В таблице обратного индекса морфологические свойства записаны в виде двоичных векторов. Поиск производится по обратному ин-

дексу с применением битовых масок, что существенно увеличивает скорость выполнения запросов к БД. Общее количество словоформ в таблице обратного индекса составляет более 46 млн.

Рис. 3. Запись данных в БД

5. Результаты и выводы

Описанный в статье прототип корпус-менеджера является оптимальным решением для задач поиска и управления данными в электронном корпусе. Функционал системы в основных функциях соответствует функционалу платформ Национального корпуса русского языка и Восточно-армянского национального корпуса, но реализация конкретных задач позволяет говорить о существенном приросте эффективности данной платформы как со стороны обширности функционала, так и со стороны скорости взаимодействия пользователя с системой. Основными преимуществами корпус-менеджера, разработанного авторами, также являются готовая поддержка татарского языка и возможность быстрой интеграции с электронными корпусами других языков, в первую очередь, тюркских, поддержка произвольных морфологических формул (рис. 4), выявление логических ошибок, открытость используемых технологий.

Для проверки пригодности, правильности, согласованности, характера изменений во времени параметров запросов, было проведено комплексное тестирование системы. Тестирование пригодности показало, что предложенные методы полностью решают поставленные задачи. Тестирование правильности, основанное

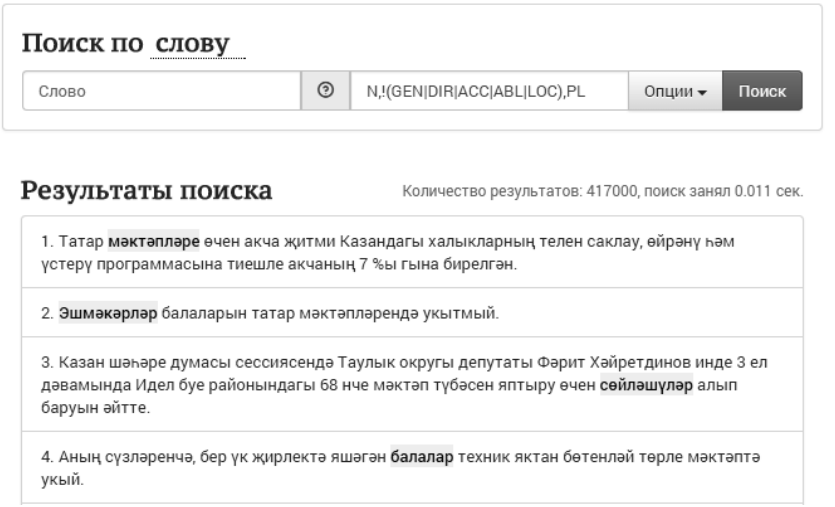


Рис. 4. Поиск с использованием произвольных морфологических формул

на сравнении с другим эталонным методом представления и обработки запросов, показало, что предложенные методы работают правильно. Тестирование согласованности и характера изменения во времени показало, что предлагаемый синтаксис лексической и морфологической составляющей поискового запроса верно интерпретируется системой, а время, необходимое для обработки и выполнения поискового запроса системой, не превышает 0,05 сек. в 98,71% случаев для лексического поиска, в 77,71% случаев для морфологического поиска и в 98,08% случаев для лексико-морфологического поиска. Во многом, таких результатов удалось добиться благодаря представлению данных в виде семантической сети и предложенным методам представления и обработки поисковых запросов.

REFERENCES

Невзорова О.А., Мухамедшин Д.Р., Билалов Р.Р. (2015а) СЕМАНТИЧЕСКИЕ АСПЕКТЫ ПРЕДСТАВЛЕНИЯ И ОБРАБОТКИ ПОИСКОВЫХ ЗАПРОСОВ В СИСТЕМЕ КОРПУС-МЕНЕДЖЕР // РЕСУРСОВ // Открытые семантические технологии проектирования интеллектуальных

систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2015): материалы II Междунар. научн.-техн. конф. (Минск, 19–21февраля 2015 г.) / редкол. : В. В. Голенков (отв. ред.) [и др.]. – Минск: БГУИР, 2015. С. 439–444.

Невзорова О.А., Мухамедшин Д.Р., Билалов Р.Р. (2015б) Корпус-менеджер для тюркских языков: основная функциональность // Труды международной конференции «Корпусная лингвистика – 2015». – СПб.: С.-Петербургский гос. Университет, филологический факультет, 2015. – С. 344–350.

Aksan, Y. et al. (2012). Construction of the Turkish National Corpus (TNC). In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012). İstanbul. Turkiye. http://www.lrec-conf.org/proceedings/lrec2012/papers.html

Dalkiliç, G., and Çebi Y., A 300 MB Turkish Corpus and Word Analysis, Advances in Information Systems (ADVIS) 2457, 2002, pp. 2002, LNCS 205–212.

Say, Bilge, Deniz Zeyrek, Kemal Oflazer and Umut Özge. “Development of a Corpus and a Treebank for Present-day Written Turkish”, (Proceedings of the Eleventh International Conference of Turkish Linguistics, August, 2002).

METU Turkish Corpus [Электронный ресурс]. URL: http://www.ii.metu.edu.tr/content/metu-turkish-corpus-access-page

Yusup Aibaidulla and Kim-Teng Lua The development of Tagged Uyghur Corpus, Proceedings of PACLIC17, 1–3 October 2003, Sentosa, Singapore, pp. 228–234.

Бускунбаева Л.А., Сиразитдинов З.А. Система разметок в национальном корпусе башкирского языка // Языки меньшинств в компьютерных технологиях: опыт, задачи и перспективы // Международная конференция. Йошкар-Ола, 2011. С. 46–51.

Шеймович, А.В. Морфологическая разметка корпуса хакасского языка // Российская тюркология. № 2(5). 2011. С. 48–61.

Салчак А.Я. Электронный корпус текстов тувинского языка. Электронный информационный журнал “Новые исследования Тувы”.– № 3. – 2012. [Электронный ресурс]. URL: http://www.tuva.asia/journal

INFERENCE RULE DISCOVERY FROM TURKISH TEXT

Gözde Gül Şahin, Eşref Adalı

Istanbul Technical University, Department of Computer Engineering, Istanbul, TURKEY

In this paper we present a framework for extraction of inference rules from Turkish documents, such as "A solved B ~ A found a solution to B". Many natural language processing tasks, such as question answering, information retrieval and machine translation can benefit tremendously from inference rules. Our framework consists of three layers: construction of dependency trees; determining predicates and their complements and finally discovery of inference rules via clustering. We use Silhouette Index (SI) of the clusters as an automatic evaluation metric.

1. Introduction

Extraction of semantic information from a structural source has advantages, however communication between people is carried out via unstructured articles, images or videos, and mostly via text. One specific goal of information extraction systems is to find inference rules, also known as variants or paraphrases, from text. Many complicated NLP tasks can benefit from extracted inference rules. Consider a query "What movies has Dustin Hoffman been in?" and a Wikipedia article with sentences "Hoffman played the character of Benjamin Braddock in Midnight Cowboy" and "Hoffman next starred in Lenny (1974)". Without knowing the relation between "X has been in movies Y", "X played a character in Y" and "X starred in Y", information retrieval system can not use the information in the article to find the answer to this question.

In literature, discovery of inference rules have been investigated for English language (Lin, 2001) and it was shown that discovered inference rules increased the performance of tasks that involve information retrieval. However, direct adaptation of these techniques to Turkish language is not possible due the language's rich morphological structure and free word order property. Moreover, to the best of our knowledge, discovery of paraphrases/inference rules has not yet been studied for Turkish. One of the reasons for the topic not being examined was the lack of annotated corpora and ready-to-use language tools such as robust and fast morphological analyzers, disambiguators

and dependency parsers. The introduction of such tools (Eryiğit, 2008) encouraged us to do this pilot study for inference rule discovery.

Our work is based on this basic assumption inspired from the studies of Lin et. al. (Lin, 2001): Even though there is no lexical similarity between words or phrases like "to play" and "to star", the sentences built up with these predicates both supply similar answers to basic questions like who, what and where. In other words, we assume that the arguments of predicates will be lexically similar. Consider the following sentences with annotated arguments:

$$[\text{Hoffman}]_{\text{Agent}} \textbf{ played } [\text{Benjamin Braddock}]_{\text{Theme}} \text{ in } [\text{Midnight Cowboy}]_{\text{Production}}$$

$$[\text{Hoffman}]_{\text{Agent}} \textbf{ starred in } [\text{Midnight Cowboy}]_{\text{Production}}$$

The predicates "play" and "star" have similar arguments like $[\text{Hoffman}]_{\text{Agent}}$ and $[\text{Midnight Cowboy}]_{\text{Production}}$, because they have similiar meanings. We assume with large enough dataset, the number of shared arguments between similiar predicates will be larger than the number of shared arguments between dissimilar predicates.

In the next section, we review previous work. Then in Section 3, we introduce our framework and go through all the layers. In Section 4, we present our evaluation method and results and finally in Section 5, we conclude.

2. Previous Studies

Schubert et. al. (Schubert, 2002) try to extract world knowledge that is obvious for human but a mystery for computers, such as "a female individual may have an arm" automatically from documents. First, sentences are syntactically parsed and predefined relations such as Subject-Object-Verb (SOV) are determined from parser output. Extracted world knowledge is filtered based on a frequency based threshold method.

Clark et. al. (Clark, 2009) focused on 12 different syntactic patterns such as SOV and Adjective-Noun (AN) and used the extracted relation to improve the parser. More than 300 relations are evaluated by 12 human judges and 70% of the relations are stated to be meaningful for humans.

Fan et. al. (Fan, 2012) presents a study that is used as a subcomponent for IBM Watson, a Jeopardy winner. They have used tools that performs

named entity recognition and coreference resolution in addition to syntactic parsers. They state that using the knowledge exracted from this subcomponent increased the success of higher level applications. Lin et. al. (Lin, 2001) used dependency trees and extracted paths with an algorithm that is based on Extended Distributional Hypothesis. To evaluate the method, candidate answers to the questions prepared for TREC-8 conference, are generated by the help of their algorithm. Then the answers are manually investigated and 32%-92% success rate is achieved.

Previous works on Turkish mostly focus on hypernym/ hyponym discovery (Yıldırım, 2012), lexicon based information extraction (Adali, 2009), (Orhan, 2011) sentence segmentation, name tagging (Tür, 2001) and topic segmentation (Tür, 2001) (Can, 2008).

3. Method

Turkish is an agglutinative language from Ural-Altaic language family. It is highly productive by means of inflectional and derivational morphology. Due to its different structure from English, previous methods explained in Section 2 can not be directly applied for Turkish. In Fig. 1, the general structure of the framework that we propose, is shown. In the first layer, dependency trees are extracted by using the method in (Eryiğit, 2008). Then, dependency trees are given as an input to the second layer, which will determine the predicates and their complements from dependency trees. This layer is also responsible for stemming of predicates and arguments in order to correctly measure lexical similarity between them. In the final layer, all binary configurations of predicate-argument structures (PAS) are calculated and clustered by using the similarity measures from (Lin, 2001).

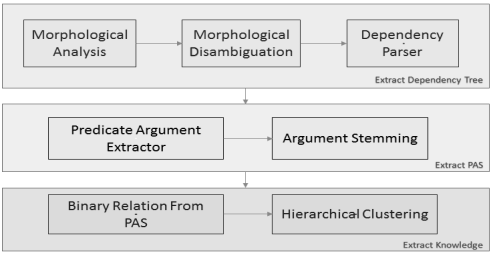


Fig. 1. Proposed Three Layered Framework

3.1. Layer 1 – Dependency Tree Extraction

For this layer, we used the data-driven dependency parsing method for Turkish proposed in (Eryiğit, 2008). An exemplary dependency tree for a Turkish sentence is shown in Fig. 2. Here, each word is splitted into inflectional groups (IG), separated by derivational boundaries (DB). Each IG is treated as if it was a separate word and annotated with its own part of speech and other necessary tags. With this representation, the links are better constructed within the sentence. Consider the word *duran* from Fig. 2. It has two IGs: The first IG, *dur (to stand)*, is a Verb, and the second IG is an adjective derived from the first IG with the morpheme *an (dur+an -> standing)*. The preceding word *şurada (there)* is linked only with the first IG *dur (to stand)* as a locative adjunct, where *kız (girl)* is linked with the second IG, *dur+an (standing)* as a modifier. Even not shown in the figure, each IG has a DERIV (derived) relation with the next IG if they are both within the boundary of the same word.

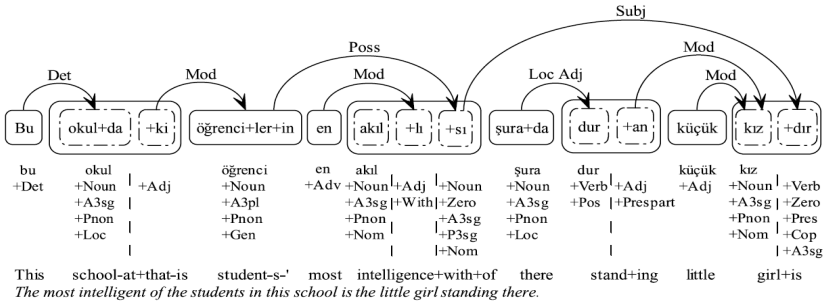


Fig. 2. Dependency tree of a Turkish sentence (taken from (Eryiğit, 2008)), where '+' sign shows morpheme boundaries, rectangle indicates word boundaries and dashed rectangle shows IG boundaries within words. Morphological analyzer results are entered below the Igs.

The essential information needed for the second layer, is mostly encoded in *ID*, *HEAD* and *DEPREL* fields of parser output as shown in Table 1. One can interpret this information as *HEAD* and *ID* being linked to each other with a *DEPREL* kind of relation. *FEATS* and *CPOSTAG* fields are also necessary for the next layer to reach the lemma of the word in argument stemming component.

Table 1

Necessary Fields from Conll Format

Field Name	Field Explanation
ID	IG counter, starting from 1 for each sentence
CPOSTAG	Coarse-grained Turkish part-of-speech tags
FEATS	Set of morphological features
HEAD	Head of the current IG
DEPREL	Dependency relation to the head

Although the output format is common for all languages, the fields CPOSTAG, FEATS and DEPREL are specific to languages.

3.2. Layer 2 – Predicate-Argument Structure Extraction

Revealing semantic information is usually achieved by identifying the complements/arguments of a predicate and assigning meaningful labels to them. Each label represents the argument's relation to its predicate and is referred as a semantic role. In the following example a sentence with annotated semantic roles are given:

$$[\text{Jess}]_{\text{Agent-Arg0}} \text{ bought } [\text{a coat}]_{\text{Theme-Arg1}} \text{ from } [\text{Abby}]_{\text{Source-Arg2}}$$

There are different conventions for semantic roles specified for languages with rich resources. One possibility is to use thematic roles defined in VerbNet (Schuler, 2006), such as Agent, Theme and Source as in the example. Another widely used annotation format is numbering arguments as in PropBank (Palmer, 2005). They both use a reference ontology to look up for the predicate and its semantic roles. Since no such ontology exists for Turkish, we only extract predicates and arguments and treat labeling task as a future work.

3.2.1. Determining Verb Predicates

In our dependency tree format, each IG is treated as if it is a separate word. This means one IG can be verb where the second IG can be noun, adjective, adverb or another verb. If IG's part-of-speech (POS) tag is Verb and the next IG's is not a Verb, then IG is considered as verb

predicate. In that case, in Fig. 2, the IGs dur and kəz-dər will be verb predicates.

Table 2

Derivational Morphology of Verbs in Turkish

Root Verb	Causative Verb
$[\text{Kız}]_{A0}$ $[\text{mont-u-nu}]_{A1}$**giy**-di.	$[\text{Oğlan}]_{A0}$ $[\text{kız-a}]_{A2}$ $[\text{mont-u-nu}]_{A1}$ **giy-dir**-di
$[\text{The girl}]_{A0}$ **put on** $[\text{her coat}]_{A1}$.	$[\text{The boy}]_{A0}$ had $[\text{the girl}]_{A2}$ **put on** $[\text{her coat}]_{A1}$
girl coat-POSS3S-ACC **put+on**	boy girl-DAT coat-POSS3S-ACC **put+on-CAUS**

In Table 2, one typical derivational morphology of Turkish verbs is shown. On the left side a Turkish root verb, giy (to put on), with annotated argument list and morphologic analysis is given. On the right side, same information for, giy-dir (to have someone put on sth.), which is causative of giy (to put on), is given. Neither the number of arguments nor the argument labels are the same. For this reason, we only consider the last IG whose POS tag is Verb, as a predicate. For example, if we have a sentence with giy-dir, only the second IG will be treated as a predicate.

3.2.2. Determining Arguments of Predicates

Arguments of predicates can follow two different paths in the dependency tree. They can either be the parent or the child of the predicate node. After the predicate node is determined by performing a Depth First Search (DFS) on dependency tree, parent and children nodes can directly be reached from that node.

In Fig. 3a, after the predicate *al (to take)*, is extracted, the parent nodes are traversed recursively till the next word boundary is reached. So, after passing the boundary of the word *al+an (sth. who took),* we reach *sözcük (word)* as the parent node of predicate *al (to take)*. Children arguments are similiarly extracted by traversing child nodes. Sentence form of the dependency tree with shared arguments shown in Fig. 3b is as follows:

Kuyumcu-dan inci al-mɘє , kɘzlar-a vermiє.
Jeweler-from pearl buy-Past , the girls-to give-Past.
*He/She bought a pearl from jeweler and gave **it** to the girls.*

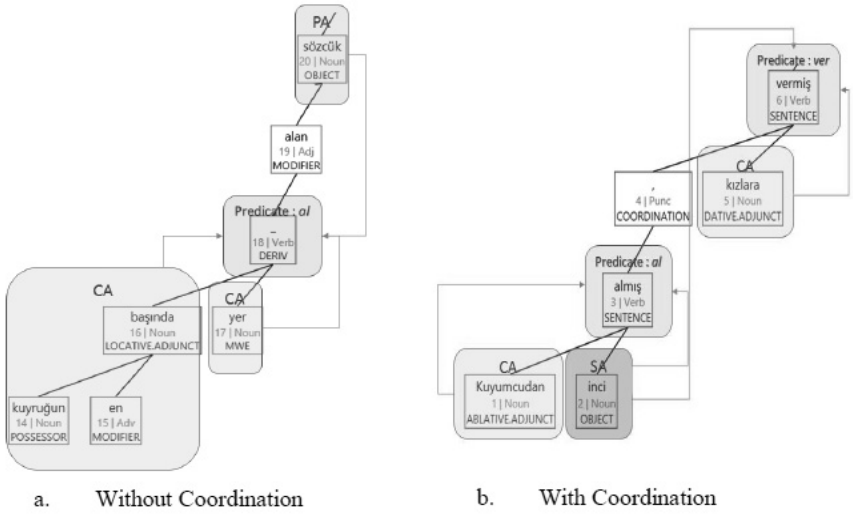


Fig. 3. Predicate – Argument relations overlaid on dependency trees.

Unlike in English, coreferences are rarely used in Turkish. In example above, the pearl that is bought and then given is, referred to as "it" as shown in bold. But in Turkish it is hard to infer it solely from syntactic structure of the sentence. In such cases, dependency parser is responsible for linking the verb predicates with a *COORDINATION* relation. Determining the shared argument takes a few extra steps, but the details will not be presented here, since it is the topic of another study.

We impose the following constraints on argument extraction:

1. Arguments with number of nodes that are higher than four, are eliminated because long paths are very rare, so not very helpful.

2. Only arguments of type Nouns are kept as in (Lin, 2001), since arguments represent values in inference rules.

Output of Turkish dependency parser supplies 24 different relation types. However only 6 of them given in Table 3, satisfy our noun constraint on arguments.

Table 3

Dependency Relations

Id	Relation Type	Example
1	Subject	Anne-m geldi. mother-my come-PAST. *My mother came.*
2	Object	Arkadaş-ı-nı yarala-dı. friend-his-ACC hurt-PAST. *He hurt his friend*.
3	Ablative.Adjunct	Annem-den gel-di-m. mother-ABL come-P1s-PAST I came from my mother.
4	Dative.Adjunct	Jack okul-a git-ti. Jack school-DAT go-PAST. Jack went to school.
5	Locative.Adjunct	Ankara'da otur-uyor-um. Ankara-LOC live-P1s-PRES. I live in Ankara.
6	Instrumental. Adjunct	Anne-si-yle gel-di. mother-his-INST come-PAST He came with his mother.

We represent predicate-argument structure (PAS) as *predicate(roleID1_arg1, roleID2_arg2,…),* where arg indicates argument and roleID represents the type of relationship between head and modifier as given in Table 3. One possible PAS extracted from Fig. 3b will be *almış(3_kuyumcudan, 2_inci),* since *Kuyumcudan* node is linked to *almış* with *Ablative.Adjunct (3)* relation, and similarly *inci* is connected to *almış* with *Object (2)* relation .

Our method relies on calculating lexical similarities between the arguments of candidate predicates. However due to rich morphology of Turkish, even lexically similar words can appear in tens of different forms such as ev+i, ev+e, ev+i+ne, evler+de etc... In order to be able to compare them, the lemma of the arguments, like ev, are found by the help of the information provided by FEATS and CPOSTAG fields of dependency nodes. After argument stemming PAS from Fig. 3b will be al(3_kuyumcu, 2_inci) and ver(2_inci, 4_kız).

3.3. Layer 3 – Predicate-Argument Structure Extraction

This final layer is responsible for converting stemmed PAS into binary relations and clustering these relations by using a similarity measure based on lexical similarities of arguments.

Binary relation between two arguments X and Y can be referred to as a semantic path. PAS can be easily turned into semantic paths by taking all possible binary configurations of its arguments. Sample paths constructed from PAS are shown in Fig 4e. Please note that constructing paths in one direction is sufficient for Turkish since the order of words are free and all necessary information is encoded in role IDs.

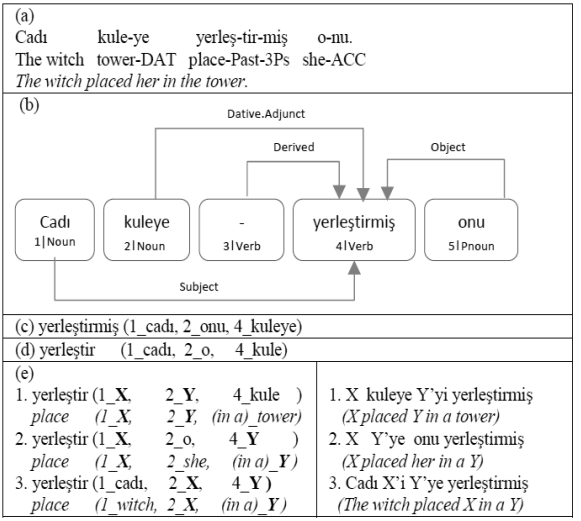


Fig. 4. All processes performed before clustering a) Input sentence taken from METU-Sabanci Corpus (Oflazer, 2005) b) Dependency parser output c) Predicate argument structure (PAS) d) Stemming PAS e) Extracting paths

The final step is clustering related paths such as "X played a character in Y" and "X starred in Y". We follow extended distributional hypothesis proposed in (Lin, 2001). According to this theorem "if two paths occur in similar context, the meanings of the paths tend to be similar". For example it is very likely to see X as an actor name such as Brad Pitt and Y as the name of the movie such as Benjamin Button, Troya etc… in both paths. In order to compute the similarity between

two paths using this theorem, we need to calculate the context of the path which corresponds to the frequency counts of all paths in a corpus and the values of X and Y for the paths. We use the same triple notation (p, SlotX, w_1) as in (Lin, 2001) where p is a path that connects two words w_1 and w_2 and SlotX indicates that w_1 appeared as X in p. As we see w_1 in X slot of p, we increase the frequency of the triple. The formula in Eq.1 can be interpreted as the similarity between paths p_1 and p_2 is actually the geometric mean of the similarities between their X and Y slots.

$$S(p_1, p_2) = \sqrt{sim(SlotX_1, SlotX_2) \times sim(SlotY_1, SlotY_2)} \quad (1)$$

Following (Lin, 2001), similarity between two slots and mutual information between a path slot and its filler are computed by Eq. 2 and 3 accordingly.

$$sim(slot_1, slot_2) = \frac{\sum_{w \in T(p_1,s) \cap T(p_2,s)} mi(p_1, s, w) + mi(p_2, s, w)}{\sum_{w \in T(p_1,s)} mi(p_1, s, w) + \sum_{w \in T(p_2,s)} mi(p_2, s, w)} \quad (2)$$

$$mi(p, Slot, w) = \log\left(\frac{|p, Slot, w| \times |*, Slot, *|}{|p, Slot, *| \times |*, Slot, w|} \right) \quad (3)$$

Basically, equations above signify that more features two paths share, the more similar they are. However not every feature is equally indicative. For example, stop words like ‘o’ (he/she) , ‘bu’ (*this*), ‘yap’ *(do)* are more frequent than words like ‘cadɔ’ *(witch)* and ‘kule’ *(tower)*. Eq. 3 used in (Lin, 2001) accounts for this problem by calculating mutual information of triples. |*symbol*| stands for the frequency of the *symbol* inside and ‘*’ is used as a regular expression.

For the languages with rich resources, it may be possible to calculate the most similar paths and evaluate them manually or use the results on a higher level application as in [4,5]. However for Turkish, there are not enough resources and higher level applications. For this reason, we choose to cluster the paths, and use automatic evaluation measures proposed for clustering techniques directly on this problem. But before clustering, the paths that contain triples *(p, SlotX, w1)* that appeared only once are eliminated.

The number of clusters are not known and it largely depends on the

size of the corpus. We expect that density inside the clusters to be similiar. In other words, we assume that when there are thousands of different ways to say a path p_1, for another p_2 extracted from the same corpus with the same techniques, there can't be only a few alternative ways to express p_2. We chose AGNES (Agglomerative Nesting), a hierarchical clustering method, because the number of clusters are not known, clusters are not so heteregenous and it can be easily implemented.

AGNES initiates each path as a seperate cluster then at each step it combines the clusters with maximum similarity. Method can continue till all relations are in the same cluster or till the number of desired clusters reached. We used Eq. 1 as the similarity measure between paths.

The distance between any two clusters, c_1 and c_2, can be taken as the minimum, average or maximum of all distances between pairs of paths p_x in c_1 and p_y in c_2. For the current problem, when merging two clusters, c_1 and c_2, we expect every single path in c_1 to be similar between all paths in c_2. Thus, we used maximum of all distances, as a distance measure between clusters.

At each iteration, we noted down the distance between clusters referred to as CD (Clustering Distance). When two clusters are similar, merging process should continue, otherwise it should stop. When CD reaches the maximum value of 1, AGNES starts to merge clusters randomly, irrelevant of their similarity. Thus, we end the clustering when CD has a value very close to 1. In Fig. 5, CD versus number of iterations for a number of 7110 extracted paths are shown. After ~60 iterations, CD value converged to 1.

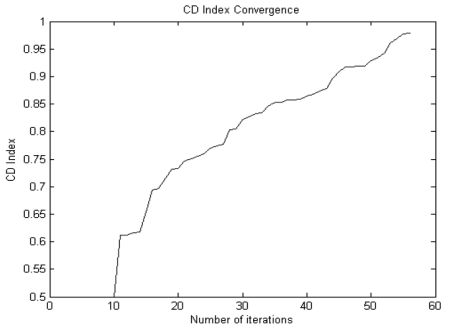


Fig. 5. Cluster Index (CI) versus number of iterations

4. Experiments and Results

We have applied our proposed framework on annotated METU-Sabanci and ITU Validation corpora (Oflazer, 2005), that are composed of ~6000 sentences together. We extracted 7110 individual paths that satisfies our constraints and grouped them into 42 clusters. Since the distribution of clusters can not be calculated, we can not evaluate homogenity of clusters and the quality of seperation by using suggested measures like Root Mean Square of Standard Deviations (RMSSTD) and R-Squared (RS). When such data is not available, the clusters can be evaluated only relying on the properties of data itself. One such method is calculting Silhouette index (SI). Consider that x_i is an extracted inference rule in cluster C_j and C_h is the closest cluster to x_i where distance measure is average-linking. Then, Silhouette index of relation x_i is given by the formula in Eq. 4. It can take values between -1 and 1, values of $s(x_i)$ closer to 1, indicates that x_i is in the right cluster. Silhoutte index of all clusters can be computed as average of SI s of all relations. We have measured our overall SI as -0.009, very close to 0. This can be interpreted as an average clustering result and room for improvement.

$$s(x_i) = \frac{d(x_i, C_h) - d(x_i, C_j)}{\max(d(x_i, C_h), d(x_i, C_j))} \tag{4}$$

In Table 4, three clusters with the highest SI scores are shown. They are obviously related to each other but necessarily has the same meaning like in cluster 26. Since the method is applied only on 6000 sentences, the number of paths in a cluster is very small, ranging between 2 and 4.

Table 4

Sample Clusters

ID	Sample Paths in Cluster
7	X'i Y'ye söylemek (Somebody tells X to Y) X'i Y'ye öğütlemek (Somebody advises Y to do X)
33	X, Y'ye gönderilmek (X is sent to Y) X, Y'ye iletilmek (X is delivered to Y)
26	5. X, Y'yi üzmek (X upsets Y) 6. X, Y'yi sakinleştirmek (X calms down Y)

7. Conclusion

In this paper, we introduced a complete framework to discover inference rules from unstructured Turkish text. To the best of our knowledge, this is the first attempt to discover such knowledge automatically from Turkish text. We introduced a method based on predicate-argument structures, relying on the fact that the paths with similiar meanings should use similar sets of words as their arguments. We proposed clustering the rules to automatically evaluate the results with Silhouette Index (SI). Our experimental results for this pilot study, showed that the framework can generate human interpretable clusters with an SI value close to 0. This work encourages us to test the framework on a larger corpus and test the results on a higher level application like an information retrieval system in the future.

REFERENCES

G. Eryiğit, J. Nivre, and K. Oflazer. Dependency Parsing of Turkish, Computational Linguistics, 34 no.3, 2008.

L. Schubert, Can we derive general world knowledge from text, in Proc. HLT, 2002, pp. 94–97.

P. Clark and P. Harrison, Large-scale extraction and use of knowledge from text, in Proc. 5th KCAP, 2009, pp. 153–160.

D. Lin and P. Pantel, Dirt - Discovery of inference rules from text, in Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2001, pp. 323–328.

J. Fan, A. Kalyanpur, D. C. Gondek, and D. A. Ferrucci, Automatic knowledge extraction from documents, IBM J. Res.& Dev., vol. 56, no. 3/4, Paper 5, pp. 5:1–5:10, May/Jul. 2012.

K. K. Schuler 2006. VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon PhD, University of Pennsylvania.

M. Palmer, P. Kingsbury, and D. Gildea. 2005. The Proposition Bank: An AnnotatedCorpus of Semantic Roles. Computational Linguistics, 31(1):71–106.

K. Oflazer, B. Say, D. Z. Hakkani-Tür, G. Tür, Building a Turkish Treebank. Invited chapter in Building and Exploiting Syntactically annotated Corpora, Anne Abeille Editor.

S. Adali, A. Sonmez, M.Gokturk. CICLing, volume 5449, 394-405. Springer, (2009).

Z. Orhan, İ. Pehlivan, V. Uslan, P. Önder. Automated Extraction of Semantic Word Relations in Turkish Lexicon. Mathematical and Computational Applications, Vol.16, No.1, pp. 13–22, 2011.

S.Yıldırım, T. Yıldız. Automatic Extraction of Turkish Hypernym-Hyponym Pairs From Large Corpus, Proceedings of COLING 2012: Demonstration Papers, pages 493–500.

G. Tür, D. Hakkani-Tür, K. Oflazer. A statistical information extraction system for Turkish, Natural Language Processing 9 (2): 181–210, 2001.

F. Can, S. Koçberber, E. Balcık, C. Kaynak, Ç. Öcalan, O. Vursavaş. Information Retrieval on Turkish Texts. Journal of the American Society for Information Science and Tech. (IST) Vol. 59, No. 3,2008, pp. 407–421.

F. Can, S. Koçberber, S. Kardaş, Ç. Öcalan and E. Uyar. New Event Detection and Topic Tracking in Turkish. Journal of the American Society for IST. Vol. 59, No. 3,2008, pp. 407–42.

THE MAIN RESULTS OF THE PROJECT OF DESIGNING TUVAN ELECTRONIC CORPUS

Aelita Salchak, Aziyana Bayir-ool

Tuvan State University
Kyzyl, Tuva, Russia

The article presents the main results of the project of designing an electronic corpus of Tuvan language. In the article short descriptions of the programs designed to work with the texts are given, «formal morphological» criteria for tagging are grounded, the models of Tuvan nominal and verbal word forms are showed.

Работа по созданию корпуса тувинского языка была начата в 2011 году при финансовой поддержке РГНФ (проект № 11-04-12073в). Исполнителями проекта были преподаватели и научные сотрудники Тувинского государственного университета: Салчак А.Я., Байыр-оол А.В., Далаа С.М., Арапчор Т.А., Хертек А.Б., Ооржак Б.Ч, Барыс-Хоо В.С., Симчит К.А., Бавуу-Сюрюн М.В. Проект выполнялся в течение 3 лет. В рамках проекта была проведена работа по переводу текстов тувинской художественной литературы, фольклорных текстов, текстов официальных документов в электронный вид, осуществлялась работа по созданию компьютерных программ по обработке текстов на тувинском языке и баз данных.

Для разметки текстов был выбран «формально-морфологический» подход.

Для обозначения грамматических признаков вводится система сокращенных грамматических помет на основе латинского алфавита и латинских названий грамматических категорий, используемых в общем языкознании. Например, если в тексте встретилась словоформа *ажылдадым* ‘я работал’ (глагол *ажылда=*, прош. время на *=ды*, 1-ое л. ед. ч. *=м*), то она будет помечена следующим образом: Rv=Past1=1sg. Для более широкого охвата количества пользователей в электронном корпусе предполагается, что поиск грамматических форм может производиться как по сокращенным грамматическим пометам на латинском языке, так и по сокращенным терминам на русском и тувинском языках: DAT / Д. п./Б.п. (дательный падеж/ бээриниң падежи).

Морфологическая разметка содержит информацию о словоизменительных признаках лексемы. Основы тувинских лексем (не-

производные и производные основы) входят в «Словарь основ тувинского языка», составленный исполнителями проекта. «Словарь основ тувинского языка» будет включен в базу данных, кроме того он будет использоваться при морфологической разметке для установления границ основ слов.

В морфологической разметке электронного корпуса тувинского языка на данном этапе работы определенными пометами обозначены грамматические признаки имени и глагола, структура словоформ и морфемы с возможными фонетическими вариантами. Образцы морфологической разметки имени и глагола представлены в виде таблиц [1, 216].

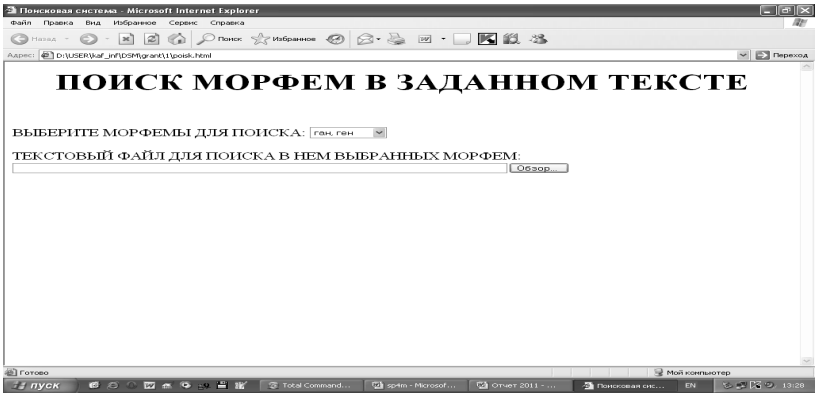
В результате выполнения проекта созданы базы данных диалектных слов, глагольных основ, частиц тувинского языка.

Созданы программы для ЭВМ: «Поиск морфем в заданном тексте»» «Поиск слов в тексте на тувинском языке», «Частотный словарь по художественным произведениям на тувинском языке», «Поиск словоформ и морфем тувинского языка».

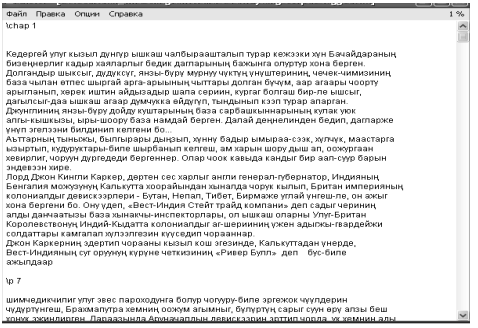
Краткая характеристика созданных программ

1. Программа «Поиск морфем в заданном тексте»» (Далаа С.М.)

Исполнителем Далаа С.М. создана программа «Поиск морфем в заданном тексте» на языке программирования JavaScript, предназначенная для поиска морфем в текстах на тувинском языке:



Данная программа работает в браузере InternetExplorer с текстами, набранными в формате txt и кодировке UTF-8. Текст заранее набирается в файле (можно в программе MicrosoftWord).



Программа успешно ищет словоформы с заданной морфемой, но возникли проблемы с омоформами. На данном этапе необходима ручная доводка соответствующей выборки.

Данная программа находится на стадии разработки и в дальнейшем она будет расширена по своим функциональным возможностям. Будет расширена база морфем глаголов, имен существительных и других частей речи, по которым можно произвести поиск в корпусе тувинского языка.

2. Программа «Частотный словарь по художественнымпроизведениям на тувинском языке» (Далаа С.М., Арапчор Т.А., Бавуу-Сюрюн М.В.)

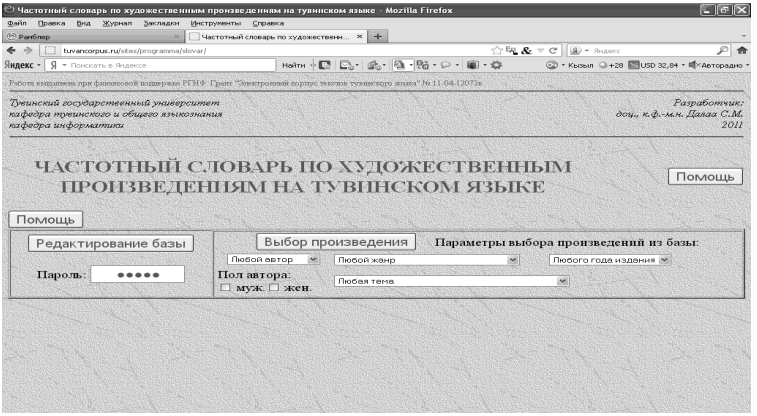
Программа «Частотный словарь по художественным произведениям на тувинском языке» на языке программирования PHP, предназначенная для создания частотных словарей в текстах на тувинском языке. Подготовкой материала занимались исполнители Салчак А.Я., Байыр-оол А.В., Барыс-Хоо В.С. и привлеченный специалист Бавуу-Сюрюн М.В.

Данная программа для ЭВМ предназначена для создания частотного словаря по художественным произведениям на тувинском языке на сайте «Электронный корпус тувинского языка» (tuvancorpus.ru) в глобальной сети Internet. Она разбивает произведения на слова в алфавитном порядке тувинского языка и подсчитывает их количество повторений в данном произведении.

Эта программа позволяет редактировать базу файлов художественных произведений на сервере (удалять и добавлять на сервер файлы). Доступ к редактированию базы возможен только по паролю.

Программа создана с помощью свободно распространяемого языка программирования PHP версии 5.2.5.

Окно при загрузке программы «Частотный словарь по художественным произведениям на тувинском языке»:



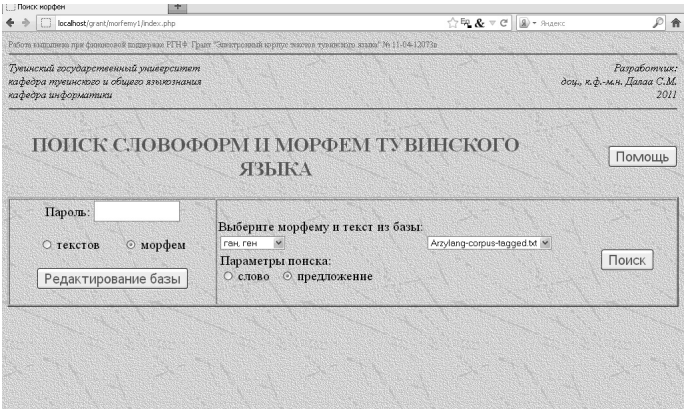
3. Программа «Поиск словоформ и морфем тувинского языка» (Далаа С. М., Арапчор Т. А., Салчак А. Я.)

Данная программа для ЭВМ предназначена для поиска создания словоформ и морфем тувинского языка в текстах на тувинском языке на сайте «Электронный корпус тувинского языка» (tuvancorpus.ru) в глобальной сети Internet. Она выводит слова или предложения с заданными морфемами из данного текста. Тексты должны быть созданы в кодировке UTF-8, что позволяет видеть буквы тувинского алфавита.

Эта программа позволяет редактировать базу текстовых файлов на тувинском языке на сервере (удалять и добавлять на сервер файлы). Кроме этого, можно изменять базу морфем тувинского языка (изменять, удалять и добавлять морфемы в базу). Доступ к базам защищен паролем.

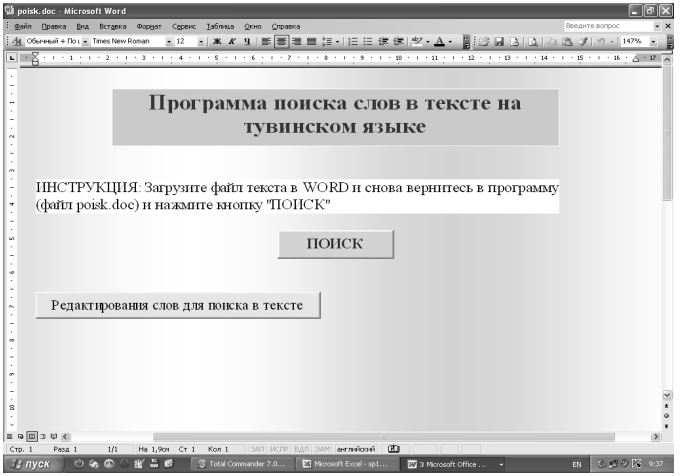
Программа создана с помощью свободно распространяемого языка программирования PHP версии 5.2.5.

Окно при загрузке программы ««Поиск словоформ и морфем тувинского языка».

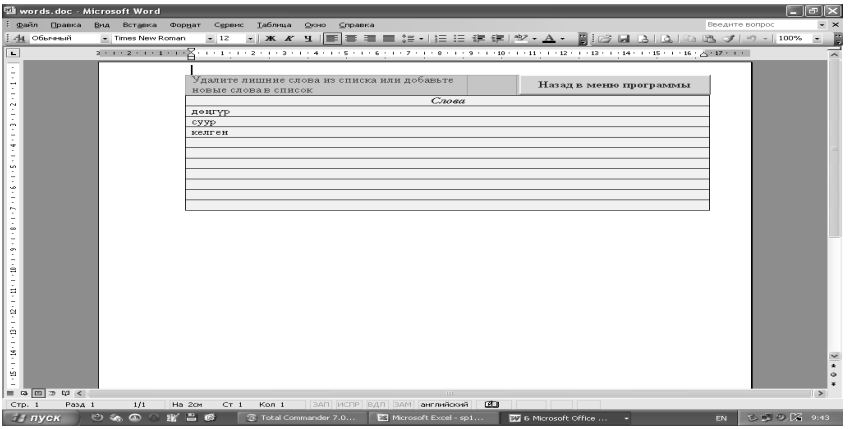


3. Поиск слов в тексте на тувинском языке (М.В. Бавуу-Сюрюн, С.М. Далаа)

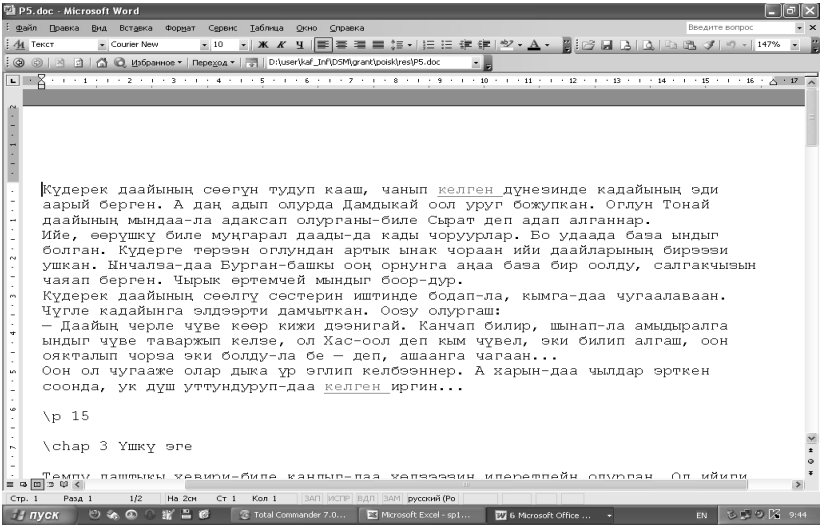
Данная программа для ЭВМ представляет собой макрос на языке программирования VisualBasicforApplication (VBA), встроенного в офисный пакет MicrosoftOffice 2003. Макрос создан в текстовом файле формата *doc*.



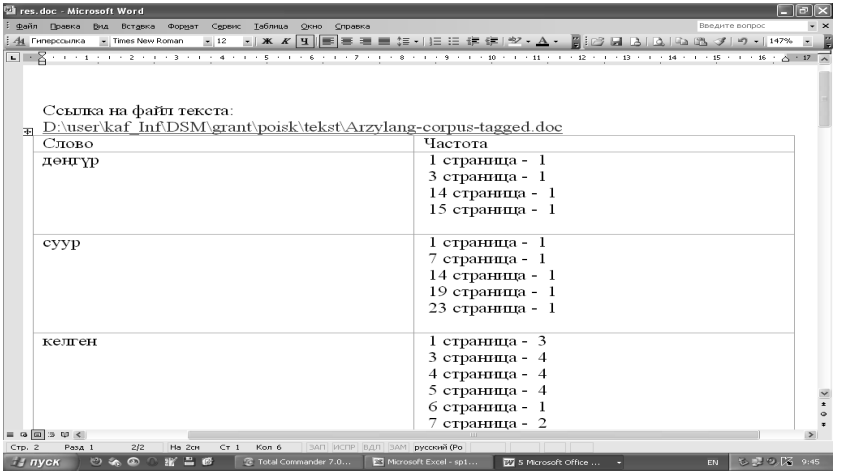
Эта программа разбивает заданный тувинский текст на страницы и находит данные слова и их словоформы на тувинском языке в этих страницах.



Результатом поиска является частота данного слова (а также его словоформ) на отдельных страницах заданного текста и все это записывается в отдельный текстовый файл формата *doc*.



Кроме этого, в файле результата создаются ссылки на страницы, по которым можно перейти и посмотреть в них найденные слова и их словоформы, выделенные красным цветом тексте.



Кроме этого, имеется текстовый файл формата *doc*, в котором можно хранить слова для поиска, и этот файл можно редактировать с помощью программы.

4. Модель тувинской словоформы (именной и глагольной)

Исполнители **Хертек А.Б.** и **Ооржак Б.Ч.** разрабатывали модель тувинской словоформы. В морфологической разметке электронного корпуса тувинского языка на данном этапе работы определенными пометами обозначены грамматические признаки имени и глагола, структура словоформ и морфемы с возможными фонетическими вариантами. Образцы морфологической разметки имени и глагола представлены в виде таблиц (см. Таблицы 1 – 4).

Таблица 1

Модель изменяемой именной словоформы и набор словоизменительных аффиксов имени

0	1	2	3	4	
Rn	**Comit**	**Num (Pl)**	**Poss**	**Case**	
				Simple declention	**Possessive declention**
	лыг	*лар*	1sg*ым*	Gen *ның*	Gen *ның*
	лиг	*лер*	1sg*ым*	Gen *ниң*	Gen *ниң*
	луг	*нар*	1sg*им*	Gen *нуң*	Gen *нуң*
	лүг	*нер*	1sg*ум*	Gen *нүң*	Gen *нүң*
	ныг	*дар*	1sg*ум*	Gen *ның*	Gen *ның*
	ниг	*дер*		Gen *ниң*	Gen *ниң*
	нуг	*тар*		Gen *нуң*	Gen *нуң*
	нүг	*тер*		Gen *нүң*	Gen *нүң*
	дыг			Gen *дың*	Gen *дың*
	диг			Gen *диң*	Gen *диң*
	дуг			Gen *дуң*	Gen *дуң*
	дүг			Gen *дүң*	Gen *дүң*
	тыг			Gen *тың*	Gen *тың*
	тиг			Gen *тиң*	Gen *тиң*
	туг			Gen *түң*	Gen *туң*
	түг			Gen *ың*	Gen *түң*
				Gen *иң*	

Таблица 2

Модель изменяемой тувинской глагольной словоформы и набор словоизменительных аффиксов глагола

0	1	2	3	4	5	6
Rv	**Rec**	**Distr**	**Perf**	**Neg**	**Tense (Pres Past,Future) Mood**	**Person (1, 2), Poss**
	ж	*гыла*	Perf*выт*	*ба*	Past 1 *ды*	1sg *м*
	ш	*гиле*	Perf*вит*	*бе*	Past 1 *ди*	2sg*ң*
		гула	Perf*вут*	*па*	Past 1 *ду*	1pl*выс*

		гүле	Perf*вут*	*пе*	Past 1 *дү*	1pl*вис*
		кыла	Perf*ывыт*	*ма*	Past 1 *ты*	1pl*вус*
		киле	Perf*ивит*	*ме*	Past 1 *ти*	1pl*вүс*
		кула	Perf*увут*		Past 1 *ту*	2pl*ңар*
		күле	Perf*үвүт*		Past 1 *тү*	2pl*ңер*
						3pl*лар*
						3pl*лер*

Таблица 3

Модель тувинского причастия и набор тувинских словоизменительных аффиксов причастия на =*ган*

0	1	2	3	4	5	6	7	
Rv	**Rec**	**Distr**	**Perf**	**Neg**	**Part**	**Person (1, 2), Poss**	**Case**	
							Simple declention	**Possessive declention**
	ж	*гыла*	Perf*выт*	*ба*	Part 1 *ган*	1sg *ым*	Gen *ның*	Gen *ның*
	ш	*гиле*	Perf*вит*	*бе*	Part 1 *ген*	1sg *им*	Gen *ниң*	Gen *ниң*
		гула	Perf*вут*	*па*	Part 1 *кан*	2sg*ың*	Gen *тың*	Gen *тың*
		гүле	Perf*вүт*	*пе*	Part 1 *кен*	2sg*иң*	Gen *тиң*	Gen *тиң*
		кыла	Perf*ывыт*	*ма*		1pl*ывыс*		
		киле	Perf*ивит*	*ме*		1pl*ивис*		
						2pl*ыңар*		
						2pl*иңер*		

Таблица 4

Модель тувинского деепричастия и набор тувинских словоизменительных аффиксов деепричастия на =*п*

0	1	2	3
Rv	**Rec**	**Distr**	**Conv**
	ж	*гыла*	Conv1 *п*
	ш	*гиле*	Conv1 *ып*
		гула	Conv1 *ип*
		гүле	Conv1 *уп*

		кыла	Conv1 *ун*
		киле	
		кула	
		куле	

Необходимые пояснения к таблицам

В таблицах последовательность следующих за корнем (основой) аффиксов в словоформе представлена пронумерованным обозначением позиций по горизонтали.

Аффиксы-показатели грамматических категорий приводятся со всеми возможными алломорфами.

Одну позицию по горизонтали могут занимать аффиксы одной или нескольких грам. категорий, каждая из которых (кроме R) может быть не выражена, например: единственное число, именительный падеж имени.

В глаголе позиции аффиксов совместно-взаимного залога (1-ая позиция Rec) и многократного вида (2-ая позиция Distr) могут меняться местами.

Условные обозначения:

Case – падеж; Conv – деепричастие; Distr – многократный вид; Future– будущее время; Mood– наклонение; Neg – отрицание; Num (Pl) – число (мн.); Part– причастие; Past – прошедшее время; Perf – законченный вид; Pers – лицо; Pres– настоящее время; Poss – принадлежность; Possessivedeclention – набор падежных аффиксов притяжательного склонения (после показателя принадлежности); Rn – основа имени; Rv – основа глагола; Rec–совместно-взаимный залог; Simpledeclention – набор падежных аффиксов простого склонения; Tense – время.

На сегодняшний день продолжается пополнение электронного корпуса тувинского языка, начата работа по семантической разметке, а также идут работы по разработке парсера и созданию поисковой системы.

ЛИТЕРАТУРА

1. Хертек А.Б., Ооржак Б.Ч. О морфологической разметке электронного корпуса текстов тувинского языка // Филологические науки. Вопросы теории и практики. Тамбов. 2012 г. № 7 (18). Ч. II. С. 214–218.

ABOUT LINGUISTIC CORPORA OF THE BASHKIR LANGUAGE

Zinnur Sirazitdinov, Liliya Buskunbaeva, Anita Ishmukhametova

Institute of History, Literature and Language of Ufa Scientific Center
of the Russian Academy of Science
Ufa, Russia

The article presents the general principles of creating corpus representing functional styles of the Bashkir language. The current state and prospects of creating new corpus the Bashkir language are given.

1. Введение

Зародившееся во второй половине XX века направление в зарубежном языкознании, связанное с компьютерной обработкой текстов больших объемов, сформировалось в новое стремительно развивающееся область филологии – корпусная лингвистика. Объектом корпусной лингвистики являются речевые материалы, реализованные как в виде письменных текстов, так и устных (фонетических) массивов данных.

Корпусы открывают перспективу для новых исследований как в области самой лингвистики, так и в смежных областях. Осознавая это, представители всех крупных языков мира разработали свои национальные корпусы. Ведутся активные корпусные разработки и по тюркским языкам: казахскому [Жұбанов 2012, Макажанов и др. 2014, Жанабекова 2014], татарскому [Сулейманов и др. 2014, Ибрагимов и др. 2014], тувинскому [Салчак идр. 2014], хакасскому [Шеймович 2011], киргизскому [Садыков и др. 2013] и др.

Учитывая актуальность корпусных разработок, башкирскими лингвистами начаты работы в данной области в четырех направлениях: 1) корпус прозаических текстов; 2) корпус публицистических текстов; 3) корпус поэтических текстов; 4) корпус фольклорных текстов [Сиразитдинов и др. 2011, Сиразитдинов и др. 2013, Сиразитдинов и др. 2014, Орехов 2014].

2. Интегрированная корпусная система

В русле указанных направлений, лабораторией лингвистики и информационных технологий ИИЯЛ УНЦ РАН разработана интегрированная система, позволяющая создавать корпусы, осущест-

влять широкий круг поисковых задач и обслуживать корпусы (редакторская работа) [Сиразитдинов, Полянин 2014]

В ходе работы над корпусными проектами лабораторией разработана интегрированная система, позволяющая создавать корпусы, осуществлять широкий круг поисковых задач и обслуживать корпусы (администрирование и сопровождение баз данных). Отметим, что сегодня многие корпусы испытывают трудности из-за отсутствия единой интегрированной системы: блок администрирования и поисковая система разделены во временном и пространственном срезе. Наша разработка, осуществленная на базе СУБД Оракл, лишена вышеназванных недостатков.

Интегрированная система состоит из двух блоков: пользовательский и администраторский.

I. Пользовательский блок включает следующие программные средства:

1. Средства определения объема корпуса, выделения пользовательского подкорпуса.

2. Поисковые средства. Программы поиска позволяют производить гибкий поиск по многим лингвистическим параметрам: словоформе, лемме, семантике, грамматическим категориям словоизменения.

3. Программы квантитативно-статистического анализа текстов корпуса. Данные средства находятся на стадии разработки. Ведутся работы по созданию подсистемы выдачи статистических распределений и их графическому представлению по любому подкорпусу, составленному пользователем. На сегодняшний день разработаны функции построения частотных словарей словоформ и лексем.

II. Блок администратора (с правами входа для сотрудников лаборатории). Включает следующие программные средства:

1. Программные средства ввода и автоматической разметки текстов. Данные средства производят морфологические и семантические разметки введенных новых текстов.

2. Средства редактирования. Предусмотрены возможности редактирования основного словаря, списков словоизменительных категорий, моделей словоизменения и правка самих текстов.

3. Средства ручного снятия грамматических и лексических неоднозначностей. Сотрудники лаборатории могут просматривать

текст по предложениям и устранять омонимичные явления, которые не разрешаются самой системой.

4. Средства принятия решений по небашкирским словам. В процессе морфологического анализа словоформ текстов система сталкивается с ситуациями, когда нет идентификации ни с основами базового словаря, ни со списком аффиксов словоизменения. Данные словоформы относятся условно к небашкирской лексике. Часть словоформ данной группы составляют опечатки, авторские неологизмы, диалектные слова и вкрапления из других языков. Программные средства позволяют исправлять опечатки, добавлять новые основы или размечать словоформы как вкрапления. Языки-источники иноязычной лексики могут добавляться или удаляться из соответствующего списка.

5. Программы статистического учета посещаемости корпуса текстов.

6. Программа экспорта любого размеченного текста из базы данных Оракл в формате xml для обмена данными с другими корпусными проектами.

3. Система морфологической разметки башкирских корпусов

Система морфологической разметки башкирских корпусов ориентирована на представление всех регулярных словоизменительных грамматических форм [Бускунбаева и др 2011, Сиразитдинов 2013]. Морфологическая информация башкирской словоформы в корпусе включает: а) частеречную характеристику; б) совокупность морфологических признаков по типу агглютинативных аффиксов словоизменения, которые подразделяются на именные и глагольные формы.

Именные морфологические признаки включают показатели 15 категорий. Глагольные морфологические признаки включают показатели 11 категорий:

В разрабатываемых нами корпусах определены следующие метаразметки:

I) Паспорт текста (для всех видов текстов)

– Автор текста.

– Название текста.

– Объем текста (указывается объем в словоформах).

– Время создания текста (если автором указано время создания, то указывается эта дата, если нет, то дата издания).

Для художественных текстов:

Тип текста: повесть, притча, рассказ, роман, сказка, триллер, эпопея, эссе и др.

Для публицистических текстов:

Тип текста: журнальный, газетный.

Тематика текста:

– политическая и социальная жизнь,
– философия,
– экономика,
– сельское хозяйство,
– искусство, культура, литература,
– наука и техника,
– образование,
– природа и путешествие,
– частная жизнь,
– спорт,
– религия,
– психология,
– медицина,
– красота и здоровье.

Жанры текстов:

– интервью, беседа,
– статья, очерк, репортаж, обозрение;
– советы,
– письма,
– обзор печати (новости из других источников),
– поздравления,
– художественные жанры,
– рецензия.

Для фольклорных текстов:

I. Эпический жанр
I.1. эпос
I.1.1. *исторический эпос*
I.1.2. *мифологический эпос*
I.1.3. *бытовой эпос*
I.2. сказки

I.2.1. *волшебные сказки*
I.2.2. *богатырские сказки*
I.2.3. *бытовые сказки*
I.2.4. *сказки о животных*
I.3. легенды
I.3.1. *мифологические*
I.3.2. *этнонимические*
I.4. предания
I.4.1. *бытовые*
I.4.1. *исторические*
II. Лирический жанр
II.1. песни
II.1.1. *исторические*
II.1.2. *лирические*
II.2. такмаки
III. лиро-эпический жанр
III.1. баиты
III.1.1. *исторические*
III.1.2. *бытовые*
III.2. мунажаты
III.3. обрядовый фольклор
IV. Афористический жанр
поговорки
пословицы
загадки
поверья
скороговорки
приметы
заклятье
проклятье
ант

4. Заключение

На данный момент автоматически размечены тексты прозаических произведений общим объемом порядка 25 миллионов словоформ и тексты публицистики объемом в 8 миллионов словоформ. Проекты этих двух корпусов запущены в сети Интернет. Над кор-

пусом фольклорных текстов работа только начата: идет сканирование текстов эпического жанра, составление базового словаря фольклора.

ЛИТЕРАТУРА

1. Жұбанов А. Қ. (2012) Қазақ тілінің аннотацияланған мәтіндер корпусындағы етесті сөздерге лексик-морфологиялық белгі-код (белгіленім) қоюдың алғышарттары/"Тілтаным", № 1, 18–25 б. (Журнал Института языкознания им. А.Байтурсынова, Казахстан, Алматы).

2. Макажанов А., О.Махамбетов, И. Сабыргалиев, Ж. Есенбаев (2014) Разработка синтаксического, лексического и морфологического наборов разметок для казахского языка / Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2014. – Казань: Издательство "ФЭН" АН РТ – С. 129–135.

3. Жанабекова А. (2014) Роль лингвистической аннотации на опыте создания национального корпуса казахского языка / Проблемы современной прикладной лингвистики. – Минск: МГЛУ. – С. 211–215.

4. Сулейманов Д.Ш., Невзорова О.А., Галиева А.М., Гатауллин А.Р., Гильмуллин Р.А., Хакимов Б.Э. (2014) Размеченный корпус татарского языка "Туган тел": аспекты реализации / Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2014. – Казань: Издательство "ФЭН" АН РТ. – С. 88–93.

5. Ибрагимов Т.И., Сайхунов М.Р. (2014) Письменный корпус татарского языка: структурные и функциональные характеристики / Актуальные проблемы диалектологии языков народов России: Материалы XIV Всероссийской научной конференции. – Уфа. – С. 261–264.

6. Салчак А.Я., Далаа С.М., Байыр-оол А.В. (2014) Электронный корпус тувинского языка: состояние, проблемы/Этногенез. История. Культура: Вторые Юсуповские чтения: Материалы Международной научной конференции, посвященной памяти Рината Мухаметовича Юсупова. – Уфа. – С. 235–238.

7. Шеймович А. В. (2011) Морфологическая разметка корпуса хакасского языка // Российская тюркология. № 2 (5). – С. 48–61.

8. Садыков Т., Шаршембаев Б. (2013) "Манас" эпосунун улттук корпусун түзүү жөнүндө // Компьютерная обработка тюркских языков.Первая международная конференция: Труды. – Астана: ЕНУ им. Л.Н. Гумилева. – 148–154 б.

9. Сиразитдинов З.А., Бускунбаева Л.А., Бардыбаева А.Д., Ишмухаметова А.Ш., Рафикова Л.Н. (2011) О разработке национального корпуса

башкирского языка / Роль и место национальной библиотеки в социокультурном пространстве: Материалы Международной научно-практической конференции. – Уфа. – С.117–126.

10. Сиразитдинов З.А., Полянин А.И., Ибрагимова А.Д., Ишмухаметова А.Ш. Корпусы башкирского языка: принципы разработки / Проблемы Востоковедения, №4, 2013, С. 65–72.

11. Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш., Ибрагимова А.Д. (2014) О создании корпуса башкирского фольклора / Урал-Алтай: через века в будущее: Материалы VI Всероссийской тюркологической конференции (с международным участием). – Уфа. – С. 86–89.

12. Орехов Б.В. (2014) Проблемы морфологической разметки башкирских текстов// Труды казанской школы по компьютерной и когнитивной лингвистике TEL-2014. – Казань. – С. 135–140.

13. Сиразитдинов З.А., Полянин А.И. (2014) Об опыте разработки интегрированной корпусной системы на базе СУДБ Оракл / Труды казанской школы по компьютерной и когнитивной лингвистике TEL-2014. – Казань. – С. 85–88.

14. Бускунбаева Л.А., Сиразитдинов З.А. (2011) Система разметок в национальном корпусе башкирского языка / Языки меньшинств в компьютерных технологиях: опыт, задачи и перспективы. – Йошкар-Ола. – С. 45–51.

15. Сиразитдинов З.А. (2013) Корпусные проекты лаборатории лингвистики и информационных технологий ИИЯЛ УНЦ РАН / Известия Уфимского научного центра РАН. № 4. С. 104–111.

TOWARDS A FREE/OPEN-SOURCE UNIVERSAL-DEPENDENCY TREEBANK FOR KAZAKH

Francis M. Tyers[a] and Jonathan Washington[b]

[a] HSL-fakultehta, UiT Norgga árktalaš universitehta,
N-9015 Tromsø, Norway
[b] Departments of Linguistics and Central Eurasian Studies, Indiana University,
Bloomington, IN 47405, USA[1] jonwashi@indiana.edu

This article describes the first steps towards a free/open-source dependency treebank for Kazakh based on universal dependency (UD) annotation standards. The treebank contains 402 sentences and is based on texts from a range of open-source and public domain sources. This ensures its free availability and extensibility. Texts in the treebank are first morphologically analysed and disambiguated and then annotated manually for dependency structure. In the article we present some issues in dependency syntax for Kazakh and how these are analysed in the universal-dependency framework. Preliminary results for statistical dependency parsing of Kazakh are reported, along with some directions for future research.

1. Introduction

This article describes work towards the development of a dependency treebank for Kazakh, a Turkic language spoken in Central Asia and Europe. Despite its status as a core Turkic language, little computational-linguistic research has been published on syntactic parsing for Kazakh. A valuable resource in the study of syntactic parsing is a treebank–a corpus of parsed text containing gold-standard syntactic annotation.

Freely available treebanks exist for many languages, such as large languages like Finnish (Haverinen et al., 2013; Voutilainen, 2011) and Polish (Woliński et al., 2011) and smaller languages like Irish (Lynn et al., 2012). To our knowledge, however, a treebank exists for only one other Turkic language, Turkish (Oflazer et al., 2003), which is unfortunately not freely available.

In building our treebank we take advantage of existing work done on tokenisation, morphological analysis and part-of-speech tagging for Kazakh. We also take a pragmatic and iterative view of development of the treebank, in line with recent work on cross-linguistic parsing with universal dependencies (De Marneffe et al., 2014).

The remainder of the paper is organised as follows. Section 2 gives some background linguistic information on Kazakh, and outlines some special challenges in parsing Kazakh. In Section 3 we describe the corpus that we annotated and the methodology used in annotating it. Section 4 gives a sketch of some decisions we have made with respect to annotation guidelines, referring back to the discussion in Section 2. For reasons of space, these guidelines are not complete, but present a subset of guidelines which are of particular interest. A small experiment in statistical dependency parsing using the corpus is presented in Section 5, and in Sections 6 and 7 we give perspectives for future work and some concluding remarks.

2. Background

2.1. Kazakh

Kazakh (қазақ тілі), a Turkic language of Central Asia and Europe, is spoken by around 13 million people in Kazakhstan, China, Mongolia, and adjacent areas (Lewis et al., 2015). While works like Балақаев et al. (1954) provide decent syntactic overviews of the language, there is little to no work on the syntax of Kazakh within modern theoretical syntactic frameworks. The authors are familiar with such work on related languages, especially Turkish (e.g., Kornfilt, 1997 and Göksel and Kerslake, 2005); while not directly consulted for this work, these works have contributed to our understanding of Kazakh syntax.

As an agglutinative language with rich morphology and agreement phenomena, Kazakh presents some interesting challenges for computational syntax. These challenges include the syntactic functions of the various "case" morphemes, problems of "zero derivation", non-finite clauses, and copulas and copula constructions. An existing morphological transducer of Kazakh (Washington et al., 2014) implements analyses of how these various phenomena occur on the morphological level. These phenomena will be described in this section, and how they were dealt with in the annotation of the treebank will be described in section 4.

In Kazakh, as in most languages with case, there is not a one-to-one relation between "case" morphemes and syntactic function (not to mention a wide range of semantic functions). The main syntactic functions of the traditionally defined cases in Kazakh are summarised in table 1.

Table 1

Primary syntactic functions of traditionally defined cases in Kazakh

Case	Morph	Functions	Examples
NOM	—	subject attributive indefinite object indefinite genitival	***Дәрігер*** *үйді көреді.* ‘**The doctor** sees the house.’ ***қонақ*** *үй* ‘hotel (lit., **guest** house)’ *Дәрігер* ***үй*** *көреді.* ‘The doctor sees **a house**.’ ***үй*** *жануарлары* ‘**house** animals’
ACC	-/NI/	definite object	*Дәрігер* ***үйді*** *көреді.* ‘The doctor sees **the house**.’
GEN	-/NIң/	definite genitival embedded subject	***үйдің*** *жануарлары* ‘the animals **of the house**’ *Ол* ***Айгүлдің*** *ойнағанына қарап тұр.* ‘She’s watching **Aygül** play.’
LOC	-/DA/	adverbial	***Үйде*** *ұйықтап жатыр.* ‘S/he’s sleeping **in the house**.’
ABL	-/DAн/	adverbial comparator	*Дәрігер* ***үйден*** *шықты.* ‘The doctor came out **of the house**.’ ***түйеден*** *үлкен* ‘bigger **than a camel**’
DAT	-/GA/	indirect object adverbial trans. causative subj.	***Ініме*** *кітап бердім.* ‘I gave **my younger brother** a book.’ ***Тойға*** *көп кісі келіпті.* ‘A lot of people came **to the feast.**’ ***Балаға*** *әнді тыңдаттым.* ‘I made **the child** listen to the song.’
INS	-/Мен/	adverbial	*Олар* ***күшікпен*** *ойнап тұр.* ‘They’re playing **with the dog**.’

As seen in the table, the morphologically unmarked nominative case (e.g., *үй* ‘the/a house’ NOM) has a wide variety of uses, including indefinite object and genitival. Definite objects are marked with the accusative case (e.g., *үйді* ‘the house’ ACC), and definite genitivals are marked with the genitive case (e.g., *үйдің* ‘of the house’ GEN). The Kazakh transducer marks all bare nominals both as ‹NOM› and ‹ATTR›, but the various functions may be disambiguated syntactically. Genitival nominals, whether definite or not, must have a corresponding nominal with possessive morphology that agrees in person and number (e.g., *үйдің есігі* ‘the/a door to the house’; *үй тапсырмасы* ‘homework’), and subject nominals must have a corresponding predicate containing e.g., a verb or a copula that agrees in person and number with the nominative nominal (e.g., *үй көшіріледі* ‘the house gets moved’). When a nominative nominal depends on another nominal that does not have possessive case, the first one must be considered attributive (e.g., *үй киім* ‘house clothes’).

The attributive use of nominals is similar to the use of adjectives, and could even be thought of as a “zero derivation” of nominals into adjectives. Interestingly, many adjectives can also be used substantively, i.e., as nominals. Table 2 shows the various functions that nominals, adjectives, and adverbs may take.

Table 2

The default (first line of each) and "zero-derived" uses of nominals, adjectives, and adverbs

Category	Function	Example
Nominals	Substantive	***Қонақтарымыз*** *бай.* '**Our guests** are rich.'
	Attributive	***қонақ*** *үй* 'hotel (lit., **guest** house)'
	Adverbial	*Оны бір* ***рет*** *көрдім.* 'I saw him/her/it one **time**.'
Adjectives	Attributive	***жақсы*** *үй* '**nice** house'
	Substantive	*ең* ***жақсы****лары* '**the nice**st **one**s'
	Adverbial	*Оны* ***жақсы*** *танимын.* 'I know him/her **well**.'
Adverbs	Adverbial	*Ол* ***кеше*** *кетті.* 'S/he left **yesterday**.'

Nominals (<N>) default to substantive and adjectives (<ADJ>) default to attributive. For the other readings, the transducer provides readings such as <ADJ><ADVL> and <N><ATTR>.

Kazakh, like most Turkic languages, makes frequent use of non-finite verb forms by deriving verbal adjectives, verbal nouns, and verbal adverbs from verbs. Verbal adjective phrases modify nouns, as in ***мені көрген*** *дәрігер* 'the doctor **that saw me**'. Verbal nouns can function as subjects or complements of verbs or copulas, as in *Дәрігер сені* ***көргенін*** *білген жоқпын.* 'I didn't know that the doctor had seen you.' and *Дәрігер сені* ***көргені*** *жақсы болыпты.* 'It's good that doctor saw you.' Verbal adverbs allow a verb phrase to function as a clausal verbal adjunct, as in *Мен* ***дәрігерді көріп*** *қуанып кеттім.* '**Seeing the doctor**, I got happy.' Verbal nouns with certain case morphology and/or occurring with certain postpositions can also function as a verbal adjunct much in the same way that verbal adverb clauses do, as in *Мен* ***дәрігерді көргенде*** *қуанып кеттім.* 'I got happy **when I saw the doctor**.', here with a verbal noun in the locative case.

Another category that may be non-overt is the copula. The primary strategy for copula constructions in Kazakh is the use of a defective verb *е-*. In the present tense, the verb itself does not surface, but agreement morphology surfaces (cliticised to the previous word) in all but the third person forms (e.g., *Мен үйде****мін***. 'I'm at home.' versus *Ол үйде.* 'S/he's at home.'). The defective copula verb also surfaces in the recent past tense (e.g., *Мен үйде еді****м***. 'I was at home.', *Інім*

үйде еді. ‘My younger brother was at home.’). Of particular interest are non-finite forms of the copula. Copula clauses can be attributive, as in ***үйі жақсы*** *дәрігер* ‘the doctor **who has a nice house**’, or literally ‘**his/her house is nice** the doctor’. While this construction never has overt copula marking, copula clauses which are complements of verbs always have an overt copula form, e.g. *Дәрігердің үйі жақсы* ***екенін*** *білген жоқпын.* ‘I didn’t know **that** the doctor’s house **was** nice.’ Here, *екен* is a suppletive verbal noun form of the copula and is marked as accusative.

How each of these issues bears on a dependency analysis will be discussed in section 4.

2.2. Treebanks

A treebank is a parsed corpus of sentences annotated syntactically following a particular syntactic theory. Two broad groups can be distinguished: phrase-structure treebanks which annotate constituency strucure, and dependency treebanks which annotate dependency structure. Some treebanks combine both.

Treebanks can be used directly for linguistic and computational linguistic research by performing search queries–for example, to extract a valency lexicon for verbs, or to study the frequency of various syntactic phenomena such as word order or nominal case usage and syntactic function.

They can also be used to train statistical parsers which can be used to annotate previously unseen texts. These parsers can be used in end-user applications such as machine translation and computer-aided language learning. According to Nivre (2008), a parser trained on a treebank of only 1,500 sentences can provide reasonable parsing accuracy.

3. Methodology

3.1. Corpus

To create the corpus, we selected a range of texts from free and public-domain sources. These texts encompassed texts from a variety of genres, such as encyclopaedic articles, folk tales, legal texts, and phrases from a phrasebook. The corpus is not entirely balanced and leans more towards encyclopaedic text. Table 3 provides a breakdown of the sources.

Table 3

Composition of the corpus. The corpus covers a range of genres and text types from free & public-domain sources

Document	Description	Sentences	Tokens	Avg. length
UN Declaration on Human Rights	Legal text on human rights	26	417	16.0
Phrasebook	Phrases from Wikitravel	38	204	5.4
Жиырма Бесінші Сөз	Philosophical text	33	467	14.2
Қожанасырдың тойға баруы	Folk tale from Wikisource	9	139	15.4
Ер Төстік	Folk tale from Wikisource	25	209	8.4
Азамат қайда?	Story for language learners	49	433	8.8
Футболдан әлем чемпионаты 2014	Wikipedia article (2014 World Cup)	12	191	15.9
Иран	Wikipedia article (Iran)	121	1714	14.2
Радиан	Wikipedia article (Radian)	2	17	8.5
Шымкент	Wikipedia article (Shymkent)	13	160	12.3
Wikipedia misc.	Random sentences from Wikipedia	74	565	7.6
		402	4516	11.2

Table 4 briefly details the various tags (i.e., syntactic functions) associated with the traditionally defined morphological cases of Kazakh as found in the corpus. Some mappings may be surprising, but will not be discussed in detail in this paper. It should also be noted that there are several known uses which are currently unattested (partitive use of ABL as DOBJ, embedded NSUBJ marked as ACC) or rare (DAT as IOBJ) in the corpus.

Table 4

The functions associated with each case in Kazakh as currently found in the corpus

	IOBJ	CMPND	NMOD	NMOD:POSS	DOBJ	ROOT	NSUBJ	VOC
NOM	—	1	199	157	72	26	279	1
ACC	—	—	—	—	102	—	—	—
GEN	—	—	120	120	—	—	12	—
DAT	1	—	113	—	13	—	—	—
LOC	—	—	113	—	—	6	—	—
ABL	—	—	64	—	—	—	—	—
INS	—	—	33	—	—	—	—	—

3.2. Preprocessing

Preprocessing the corpus consists of running the text through the Kazakh morphological analyser (Washington et al., 2014), which also performs tokenisation of multiword units based the longest match left-

to-right. Tokeni-sation for Kazakh is a non-trivial task, and so we do not simply take space as a delimiter. The morphological analyser returns all the possible morphological analyses for each word based on a lexicon of around 20,000 lexemes. After tokenisation and morphological analysis, the text is processed with a constraint-grammar based disambiguator for Kazakh consisting of 113 rules which remove inappropriate analyses in context. This reduces the average number of analyses per word from around 3.4 to around 1.7.

3.2.1. Tokens and words

Tokenisation of the corpus is performed by our morphological analyser. This analyser performs tokenisation on the basis of a left-to-right longest match algorithm described in Garrido-Alenda et al. (2002). Simple tokens such as *толқулар* 'riots' are maintained as a single token, and their lemma and morphological analysis is returned. Multiword units such as *ауа райы* 'weather' and *ата-анасының* 'of their parents' are combined into a single token. Abbreviations and numerals which bear case, such as *АҚШ-пен* 'with the USA' and *90%-ына* 'to 90%' are analysed as a single token, as are light verb constructions such as *пайда бол* 'to appear' and tense forms written with space like *оқыған жоқ*, the third-person negative past of *оқы-* 'to study/read'.

In some cases a single token is split into two tokens, as with the aorist copula suffixes, e.g., *үйдемін* 'I am in the house' is tokenised as *үй*.LOC + *е*.cop.aor.sg1. Furthermore, two input tokens may result in three output tokens, e.g., *бар ма?* 'is there?' is tokenised as *бар*.ADJ + *е*.cop.aor.sg3 + *ма*.QST.

4. Annotation guidelines

4.1. Copula

The copula (both *е-* and *бол-)* is a challenging problem for dependency analysis of Kazakh. The universal dependency guidelines state that the copula should be the dependent of the lexical predicate. However, in many cases the copula in Kazakh is found in embedded clauses, which morphologically acts much more like the head of the embedded clause.

We have uniformly annotated the copula as a leaf node with the predicate, or adverbial as the head of the structure. For certain structures this is convenient, such as the bare copula in phrases like Ia or Ib, but

for phrases where the copula is part of an embedded clause this is not necessarily the most effective choice. In Ic, the copula holds all the morphological information, including agreement with the subject and accusative marking for the embedded clause.

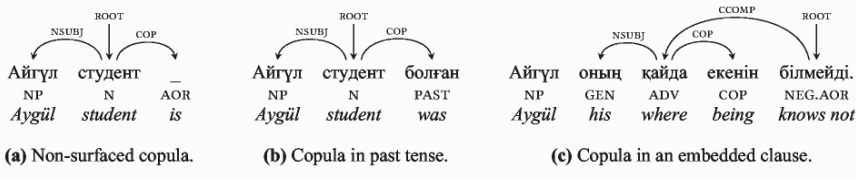


Fig. 1. Dependency trees of copula constructions.

4.2. Coordination

One difference in our annotation scheme compared to the standard universal dependency analysis is with coordination. While the universal dependency scheme takes the first conjunct as the head, we take the last. This decision was made based on the fact that Kazakh is a head-final language and morphological marking is only obligatory on the last conjunct in a series. Furthermore, experiments in representing coordination in other predominantly head-final languages have found that the final-conjunct head analysis results in better parser accuracy (Bengoetxea and Gojenola, 2009). Figure 2 shows an example of our coordination strategy.

4.3. Complex nominals

There are different ways in which two nominals may occur together to act as a single nominal. Compounds are formed by an attributive nominal (morphologically indistinguishable from the bare / nominative form, but tagged with <ATTR>) preceding another nominal, as shown in 3a. An indefinite genitive construction is formed by an indefinite genitive

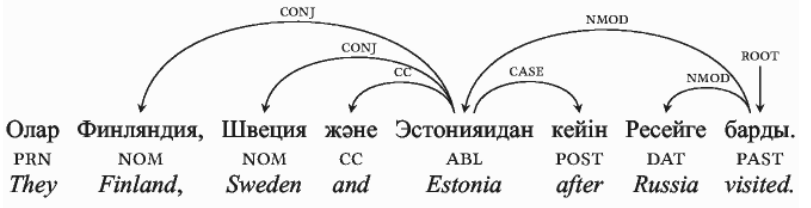
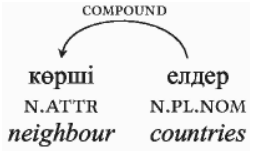
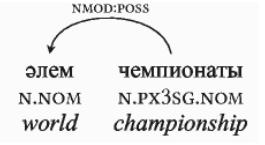


Fig. 2. Coordination: All conjucts are attached to the final conjunct, which is the head of the coordinated phrase

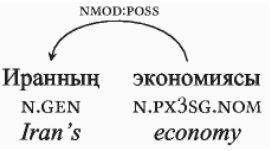
nominal (morphologically indistinguishable from the bare / nominative form, and tagged with <NOM>) preceding a nominal that has third-person possessive morphology, as shown in 3b. A definite genitive construction is formed by a genitive-marked nominal (tagged <GEN>) precding a nominal that has third-person possessive morphology, as shown in 3c.



(a) Compounding:an attributive nominal depending on a nominal



(b) Indefinite genitive: An indefinite genitive depending on a nominal.



(c) Definite genitive: A definite genitive depending on a nominal

Fig. 3. Dependency trees of complex nominal relations

As seen in the graphs, the compound relationship of an attributive nominal depending on another nominal is labelled COMPOUND, and genitive relations are considered NMOD:POSS, regardless of whether there is a definite or indefinite genitive construction.

4.4. Non-finite clauses

As discussed in section 2.1, Kazakh makes extensive use of non-finite clauses, including verbal adjective clauses, verbal noun clauses, and verbal adverb clauses.

Verbal adjective clauses modify a head nominal, effectively allowing a whole verb phrase to act as an adjective. The dependency relation between them is ACL, per UD documentation. An example is provided in figure 4.

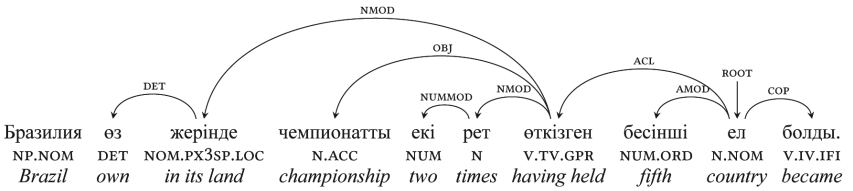


Fig. 4. Verbal adjectives: verbal adjectives are the head of everything in their own clause and are an ACL dependency of the noun they modify

Verbal adverb clauses in Kazakh act as a clausal verbal adjunct, essentially allowing a whole verb phrase to act as an adverb. The dependency relation between the verbal adverb and the head verb is ADVCL, per UD documentation. An example is provided in figure 5.

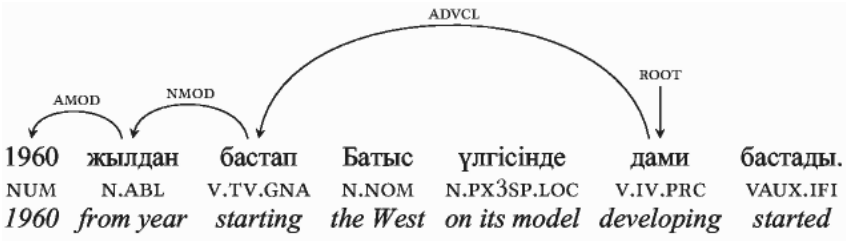


Fig. 5. Verbal adverbs: verbal adverbs are the head of everything in their own clause and are an ADVCL dependency of the verb they are subordinate to

Verbal nouns allow a whole verb phrase to be used as a nominal, and the resulting verbal noun phrase may be used as a subject or object of a verb. When used as the subject of a verb, verbal noun phrases receive the UD label NSUBJ. When used as the object of a verb, verbal noun phrases receive the UD label CCOMP, unless the subject of the subordinate clause is obligatorily identical to the subject of the main clause, in which case it receives XCOMP. An example of a verbal noun phrase used as a CCOMP dependency of the main verb is given in figure 6.

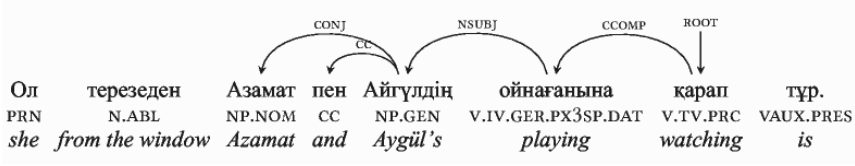


Fig. 6. Verbal nouns: verbal nouns are the head of everything in their own clause and can be a CCOMP dependency of the verb they are the complement of. Note here that the subject noun phrase of the subordinate clause is in genitive case

As mentioned earlier, verbal nouns combined with certain case morphology and/or postpositions can function as an adverbial adjunct to a main clause, similar to verbal adverb clauses. In such instances, they receive the UD label ADVCL, and example of which is given in figure 7.

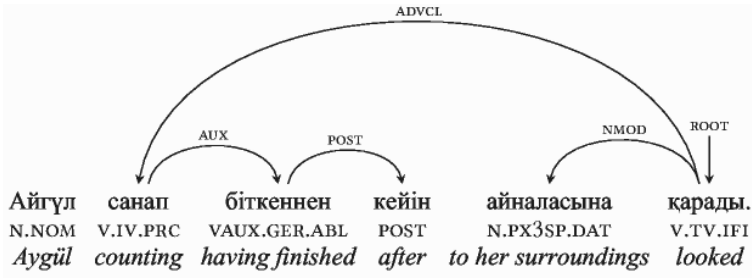


Fig. 7. Verbal nouns with case/postpositions: verbal nouns with certain case morphology and/or postpositions are the head of everything in their own clause and can be an ADVCL dependency of the verb they are the complement of

5. Evaluation

In order to test the utility of the treebank in a real-world setting, we trained and evaluated a number of models using the popular MaltParser tool (Nivre et al., 2007). MaltParser is a toolkit for data-driven dependency parsing; it can learn a parsing model from treebank data and apply this model to parse unseen sentences. The parser has a large number of options and parameters that need to be optimised. To select the best parser configuration we relied on MaltOptimiser (Ballesteros and Nivre, 2015). The optimiser was run separately for each of the model configurations, and the best combination of parameters was selected.

As the treebank takes advantage of the new tokenisation standards in the CoNLL-U format, and MaltParser only supports CoNLL-X, certain transformations were needed to perform the experiments. The corpus was flattened with conjoined tokens receiving a dummy surface form. The converted corpus is available alongside the original.

Table 5

Preliminary parsing results from MaltParser using different models. The numbers in brackets denote the upper and lower bounds found during cross-validation

Features	Algorithm	LAS [range]	UAS [range]
surface	nivreeager	42.8 [38.9, 45.9]	58.4 [52.4, 62.9]
surface+lemma	nivreeager	42.8 [38.9, 45.9]	58.4 [52.4, 62.9]
surface+lemma+POS	nivreeager	64.9 [66.4, 67.1]	77.0 [76.8, 80.0]
surface+lemma+POS+MSD	nivreeager	76.8 [70.9, 80.0]	81.4 [75.7, 85.2]

[1] http://svn.code.sf.net/p/apertium/svn/languages/apertium-kaz/texts/puupankki/

To perform 10-fold cross-validation we randomised the order of sentences in the corpus and split it into 10 equally-sized parts. In each iteration we held out one part for testing and used the rest for training. We calculated the labelled-attachment score (LAS) and unlabelled-attachment score (UAS) for each of the models.

The results we obtain, shown in table 5, are similar to those obtained with similar sized treebanks, for example the Irish treebank of Lynn et al. (2012), for which an LAS of 63.3 and a UAS of 73.3 were reported with the best model. Adding structural features to the model substantially improves the performance of the parser.

6. Future work

Future work will focus on improving the annotation guidelines and the consistency of annotation in the corpus. We will also study the possibility of deepening the annotation with Turkic-specific relations. When we have stable annotation guidelines we intend to extend the corpus with more texts. We would also like to work on cross-lingual dependency parsing, that is, applying a model learnt on the Kazakh treebank to other Turkic languages such as Tatar, Kumyk, and Tuvan. We have morphological analysers for these languages which have compatible tagsets for morphological features and as such it should be possible to learn a delexicalised model based on these features. As Turkic syntax is broadly homogenous, this presents a promising avenue for future work.

7. Concluding remarks

We have presented the first steps towards a free/open-source dependency treebank for Kazakh with annotation based on the universal dependencies. The treebank is small, but provides a base for bootstrapping further. Performance of a state-of-the-art statistical parser trained on the treebank is comparable to other treebanks of similar size.

Acknowledgements

The authors would like to acknowledge Tolgonay Kubatova; Zhenisbek Assylbekov, Aida Sundetova, and colleagues; and Aibek Makazhanov for the various ways in which they each contributed to this research.

REFERENCES

Ballesteros, Miguel and Joakim Nivre (2015). "MaltOptimizer: Fast and effective parser optimization". In: *Natural Language Engineering* FirstView, pp. 1–27.

Bengoetxea, Kepa and Koldo Gojenola (2009). "Exploring Treebank Transformations in Dependency Parsing". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP),* pp. 33–38.

De Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning (2014). "Universal Stanford Dependencies: a Cross-Linguistic Typology". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by N. Calzolari, Kh. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis. Reykjavik, Iceland.

Garrido-Alenda, Alicia, Mikel L. Forcada, and Rafael C. Carrasco (2002). "Incremental construction and maintenance of morphological analysers based on augmented letter transducers". In: *Proceedings of the Conference on heoretical and Methodological Issues in Machine Translation,* pp. 53–62.

Goksel, Ash and Celia Kerslake (2005). *Turkish: A Comprehensive Grammar.* London: Routledge.

Haverinen, Katri, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter (2013). "Building the essential resources for Finnish: the Turku Dependency Treebank". In: *Language Resources and Evaluation.* In press. Available online., pp. 1–39.

Kornfilt, Jaklin (1997). *Turkish.* London: Routledge.

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig, eds. (2015). *Ethnologue: Languages of the World.* Eighteenth. Dallas, Texas: SIL International.

Lynn, Teresa, Özlem Çentinoğlu, Jennifer Foster, Elaine Uí Dhonnchadha, Mark Dras, and Josef van Genabith (2012). "Irish Treebanking and Parsing: A Preliminary Evaluation". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by N. Calzolari, Kh. T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis

Nivre, J., J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kubler, S. Marinov, and E. Marsi (2007). "MaltParser: A language-independent system for data-driven dependency parsing". In: *Natural Language Engineering* 13.2, pp. 95–135.

Nivre, Joakim (2008). "Algorithms for deterministic incremental dependency parsing". In: *Computational Linguistics* 34, pp. 513–553.

Oflazer, Kemal, Bilge Say, Hakkani-Tür, Dilek Zeynep, and Gökhan Tür (2003). “Building a Turkish Treebank”. English. In: *Treebanks*. Ed. by Anne Abeillé. Vol. 20. Text, Speech and Language Technology. Springer Netherlands, pp. 261–277.

Voutilainen, Atro (2011). “FinnTreeBank: Creating a research resource and service for language researchers with Constraint Grammar”. In: *Proceedings oftheNODALIDA 2011 workshop Constraint Grammar Applications*.

Washington, Jonathan, Ilnar Salimzyanov, and Francis Tyers (2014). “Finite-State Morphological Transducers for Three Kypchak Languages”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*. Ed. by N. Calzolari, Kh. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis. Reykjavik, Iceland.

Woliński, Marcin, Katarzyna Głowińska, and Marek Świdziński (2011). “A Preliminary Version of Składnica–a Treebank of Polish”. In: *Proceedings of the 5th Language & Technology Conference*. Ed. by Zygmunt Vetulani. Poznań, Poland, pp. 299–303.

Балақаев, М., А. Исқақов, С. Кеңесбаев, Ғ. Мұсабаев, and Н. Сауранбаев (1954). *Қазіргі қазақ тілі : лексика, фонетика, грамматика*. Алматы: Қазақ ССР Ғылым Академиясы.

SECTION 4 | **UNITURK SEMINAR**

MORPHOLOGICAL DISAMBIGUATION IN CORPUS OF TATAR LANGUAGE

Ramil Gataullin, Rinat Gilmullin

Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan, Russia
Kazan Federal university, Kazan, Russia

This paper provides an overview of methods for morphological disambiguation and analysis of their applicability to the Tatar language. Since morphological disambiguation is part of word sense disambiguation, it becomes one of important phases of natural language processing. Research on this problem has been conducted since 1960s, so there are several approaches to disambiguation. In the beginning, contextual rules methods were used, whereas later statistical methods were employedfor this task. Most of these approaches are language independent, and for some languages (e.g. English)the accuracy reaches 97%. According to our research, contextual rules methods are applicable to the Tatar language as well, but thorough linguistic analysis of each morphological ambiguity type(of more than 7000) is necessary. Additionally, these are not all possible types of morphological ambiguity, only the most common ones. Due to the agglutinative nature of the Tatar language, the amount of them is theoretically infinite.

Statistical methods are divided into two groups: with supervised learning and with unsupervised learning. Models with supervised learning use a fully annotated and disambiguated part of a language corpus. By this time, researchers of the Research Institute of Applied Semiotics have prepared such part of the corpus using their own crowdsourcing web application. The second group, which employs supervised learning, requires only an annotated corpus part without disambiguated cases. Their further applicability to the Tatar language requires investigation.

1. Введение

Морфологический анализ текста является одним из основных этапов предварительной обработки при решении большинства задач автоматической обработки текстов. Если задача снятия лексической многозначности (англ, word sense disambiguation) представляет собой установление значений слов или составных терминов в соответствии с контекстом, то задача снятия морфологической (грамматической) многозначности ставит целью определение части речи слов (англ, part of speech tagging) [Турдаков 2010]. Современные системы способны решать эту задачу, показывая для некоторых языков точность более 97%. Но в некоторых случаях, сложность задачи решения морфологической многозначности сопоставима со сложностью задачи решения лексической многозначности.

Как показывает анализ существующих методов, проблема снятия морфологической многозначности решалась исследователями разными способами. Первые алгоритмы были основаны на правилах. Позже для решения задачи разрешения многозначности начали применяться статистические алгоритмы. Один из таких методов, основанный на модели Маркова, уже считается классическим [Бобичев 2007]. Многие методы не зависят от языка. Однако у каждого языка есть свои особенности, которые обязательно должны учитываться при анализе.

Принадлежность к агглютинативному типу определяет особенности морфологического строения слова в татарском языке: словоформы формируются путем присоединения к основе словообразовательных и словоизменительных аффиксов, каждое из которых выражает отдельное грамматическое значение. Принимая во внимание эти особенности, для татарского языка был разработан морфологический анализатор на базе двухуровневой модели морфологии татарского языка [Гильмуллин 2009].

Морфологические характеристики относительно легко распознаются при автоматическом анализе аффиксального состава словоформы. Задача же выявления и автоматической разметки (лексико-грамматических разрядов) различных частей речи не может быть решена с учетом лишь морфологических данных [Сулейманов 2011].

Согласно анализу создаваемого электронного корпуса татарского языка, в основном к проблеме разрешения морфологической многозначности в татарском языке относится проблема разрешения функциональной омонимии. Функциональная омонимия – омонимия, когда слова совпадают в написании лишь в определенных формах, являясь при этом разными частями речи [Хакимов и др. 2015].

Возвращаясь к методам, и принимая во внимание данные статистических исследований корпуса, можно предположить применимость для модели морфологий и татарского языка как контекстных, так и статистических методов.

Метод контекстных правил рассматривался автором в дипломной работе, итогом которой являлась разработка программного инструментария для разрешения морфологической многозначности [Гатауллин и др. 2014]. Данный метод оказался вполне работоспособным. Вместе с тем необходимо осуществить дальнейшую тщательную лингвистическую экспертизу каждого типа омонимии при заполнении контекстных правил.

Что касается статистико-вероятностного метода, то данные методы делятся на модели, обучающиеся с учителем, и без учителя. Для первых моделей для правильной работы необходим размеченный корпус со снятой многозначностью. Эта задача в настоящее время решается специалистами института прикладной семиотики Академии наук РТ, разработан веб-интерфейс для ручной разметки корпуса и осуществляется этап ручного снятия морфологической многозначности в выделенной части корпуса. Для вторых, т.е. для моделей с обучением без учителя, нужны только корпусные данные, без разрешенных случаев многозначных разборов. Применимость данных моделей для татарского языка до конца еще не изучена.

2. Метод контекстных правил

Теоретические исследования по проблеме разрешения омонимии в текстах имеют многолетнюю историю. Еще в конце 1950-х годов в работах [Harper 1956] основным способом снятия омонимии признавалось изучение и описание тех контекстных условий, в которых реализуется то или иное значение слова.

Актуальным для исследователей являлся вопрос о минимальном разрешающем контексте. В этой связи заслуживают упомина-

ния результаты, полученные [Caplan 1955] по исследованию минимального разрешающего контекста. Общие выводы A. Caplan сводятся к тому, что наиболее практичным является контекст, состоящий из одного слова слева и одного слова справа от анализируемой многозначной лексемы. Если же одно из слов окружения – «particle», то следует «усилить» контекст до двух слов с обеих сторон [Caplan 1955].

Исследования такого подхода для русского языка показали, что его применимость в реальных контекстах вряд ли возможна. Реальная ситуация с разрешением лексической омонимии в русском языке значительно сложнее и не может быть разрешена на основе упрощенных схем. Для решения этой проблемы для русского языка была предложена усложненная структура правил, а также предполагается в качестве контекста использовать все предложение [Зинькина и др. 2005].

Традиционно в центре внимания исследователей омонимии в татарском языке были лексические омонимы. Определенная строгость агглютинативной синтаксической структуры позволяет рассчитывать на обнаружение четких контекстных ограничений. Однако подход, основанный на правилах, как показали также наши исследования, является чрезвычайно трудоемким, требует проведения тщательной лингвистической экспертизы каждого типа омонимии. Абсолютно точное разрешение омонимии на основе метода контекстных правил также не представляется возможным по многим причинам. [Гатауллин и др. 2014]

Для каждого типа функциональной омонимии разрабатывается обобщенное правило разрешения омонимии данного типа. Обобщенное правило представляет собой упорядоченную совокупность правил, записанных на специальном формальном языке. Каждое правило внутри совокупности фиксирует некоторый разрешающий контекст. Структура задает порядок применения правил, который базируется на оценке частотности контекстов.

Рассмотрим разрешение функциональной омонимии типа (V+Refl)/(N+3PossSg+Acc), где (V+Refl) – глагол с аффиксом возвратно-страдательного залога, (N+3PossSg+Acc) – существительное с аффиксами принадлежности 3 лица единственного числа, на примере омоформы: асылын.

Варианты аффиксальной структуры омоформы:

ас(V)+ЫЛ(Refl)+ЫН(Refl) «повисьнуть» (в предложений «Муенга асылын»)

асыл(N)+СЫ(3PossSg)+нЫ(Acc) «суть (чего, кого-л.)» («Гомернең асылын аңлагыз»)

Потенциальные модели, главные компоненты и семантика словосочетаний:

в качестве зависимого компонента не встречается;

N+Acc→V (главный компонент – глагол, семантика прямого объекта).

Данный тип аффиксальной омонимии разрешается следующим правилом:

если в правом контексте находится глагол, возможна потенциальная модель словосочетания N+Acc→V (2).

Соответственно, если реализуется данная модель словосочетания, омонимия разрешается по 2-му варианту морфемной структуры, т.е. N+3PossSg+Acc: асыл(N)+СЫ(3PossSg)+нЫ(Acc).

В формализованном виде правило будет выглядеть так [Гатауллин и др. 2014]:

$$if\ (X_{1 \cap Acc} V^*)\ then\ X = N + 3PossSg + Acc$$
$$Else X = V + Refl$$

3. Статистико-вероятностные методы

В тех же 1950–1960х годах, вслед за контекстными методами стали появляться статистико-вероятностные методы. Но из-за того, что для работы им требуются большие информационные ресурсы, такие как размеченные корпусы текстов, детальные словари, то эксперименты и дальнейшее применение было оставлено до лучших времен. При появлении более или менее репрезентабельных электронных корпусов, эксперименты с этими методами показали достаточно хорошие результаты. Для английского языка, как для языка с бедной морфологией, задача снятия морфологической омонимии сводится, как правило, к проблеме разрешения многозначности на уровне частей речи (так называемого POS-теггинга). Такие алгоритмы работают достаточно хорошо и обычно демонстрируют не менее 96% точности [Турдаков 2010].

Статистические методы для разрешения морфологической омонимии применительно к русскому языку стали использовать срав-

нительно недавно. При адаптации некоторых методов для русского языка, нужно учесть некоторые особенности языка. Во-первых, морфологическая омонимия в русском языке, не сводится к омонимии частеречной, а охватывает большое количество различных грамматических признаков. Во-вторых, хорошая работа статистических моделей на материале английских текстов объясняется тем, что в английском языке фиксированный порядок слов. В русском языке, напротив, порядок слов свободный, так что предполагается, что количество возможных контекстов из-за этого увеличивается и эффективность обучения простой модели, основанной на локальных зависимостях, снижается. Поэтому, наряду с марковскими моделями, для снятия морфологической омонимии в русском языке используются более сложные статистические модели (ср., например, Петроченков) или гибридные системы, в которых статистика дополняется набором правил [Лакомкин и др. 2013].

Как отмечалось ранее, ввиду того, что корпус татарского языка только разрабатывается, экспериментальная проверка применимости статистико-вероятностных методов для татарского языка на данный момент не представляется возможным. Разрабатываемый корпус находится на этапе морфологической разметки и ручного снятия морфологических многозначностей. Морфологическая разметка осуществляется автоматически адаптированным морфологическим анализатором на базе двухуровневой модели морфологии татарского языка. А ручное снятие морфологической многозначности с корпуса осуществляется посредством веб-интерфейса, описанном в работе [Гатауллин и др. 2015]

4. Заключение

В этой работе представлен аналитический обзор основных методов разрешения морфологической многозначности и их применимость для татарского языка. Точность работы анализированных методов составляет не ниже 95%. В основном методы языконезависимы, но ввиду того, что многие методы строились для английского языка, морфология которого не очень сложна, а также имея ввиду, что высокие точности анализа получены для английских текстов, вопрос применимости и эффективности для татарского языка требует дополнительных исследований.

Необходимость высокой точности (не менее 95%) при разрешении морфологической многозначности обусловливается необходимостью результатов этого анализа в разрешении лексической многозначности. Задача же разрешения лексической многозначности является следующим шагом в развитии компьютерной лингвистики для татарского языка.

ЛИТЕРАТУРА

1. Бобичев В.Л. (2007), Автоматическое снятие морфологической многозначности при разметке корпуса. М.

2. Гатауллин Р.Р., Сулейманов Д.Ш., Гильмуллин Р.А. (2014), Программный инструментарий для разрешения морфологической многозначности в татарском языке. Открытые семантические технологии проектирования интеллектуальных систем «OSTIS–2014»: материалы IV международной научно-технической конференции (Минск, 20–22 февраля 2014 года), с. 503–508.

3. Гатауллин Р.Р., Гильмуллин Р.А. (2015), Веб-интейрфейс для снятия морфологической многозначности в корпусе татарского языка. Открытые семантические технологии проектирования интеллектуальных систем «OSTIS–2014»: материалы IV международной научно-технической конференции (Минск, 19–21 февраля 2015 года), с. 451–454.

4. Гильмуллин Р.А. (2009), Разработка файла морфотактических правил для глагольных групп татарского языка. Материалы VII Международного симпозиума «Языковые контакты Поволжья», Казань.

5. Зинькина Ю.В., Пяткин Н.В., Невзорова О.А. (2005), Разрешение функциональной омонимии в русском языке на основе контекстных правил. Труды межд. конф. «Диалог'2005». М. с. 198–202.

6. Лакомкин Е.Д., Пузыревский Е.Д., Рыжова Д.А. (2013), Анализ статистических алгоритмов снятия морфологической омонимии в русском языке. НИУВШЭ, компания Яндекс, М.

7. Сулейманов Д.Ш., Хакимов Б.Э., Гильмуллин Р.А. (2011), Корпус татарского языка: концептуальные и лингвистические аспекты. ВЕСТНИК ТГГПУ. Казань.

8. Турдаков Д.Ю. (2010), Методы и программные средства разрешения лексической многозначности терминов на основе сетей документов. М.

9. Хакимов Б.Э., Гильмуллин Р.А., Гатауллин Р.Р. (2015), Разрешение грамматической многозначности в корпусе татарского языка. Ученые записки Казанского университета. Казань.

MORPHOLOGICAL STANDARD OF CHUVASH CORPUS: INFORMATION ON MORPHOLOGICAL CHARACTERISTICS AND ARCHITECTURE OF GRAMMAR DICTIONARY [1]

Aleksey Gubanov

Chuvash State Institute of Humanities
Cheboksary, Chuvashia, Russia

The article describes the models and methods of morphological standards of the National Corpus of the Chuvash language, information about the typological characteristics of the Chuvash language is also given. The description of architecture of Grammatical dictionary of the Chuvash language, which could be used in multilingual computing applications, is presented.

Создание Национальных корпусов языков, развитие систем автоматической обработки естественного языка ставят задачу разработки определенных стандартов представления грамматических данных.

В связи с этим научно-практическим семинаром "Унификация систем грамматической разметки в корпусах тюркских языков (семинар UniTurk)" [8] принята своевременная резолюция, где говорится, что «создание электронных лингвистических корпусов выдвигает перед разработчиками широкий спектр проблем и задач, успешное решение которых требует соединения результатов лингвистических исследований и современных компьютерных методов анализа языковых данных. Возможности корпуса во многом определяет система аннотации (разметки)». Действительно, актуальной задачей является разработка определенных стандартов представления данных: стандартизация форматов и стандартизация концепций. Морфологический стандарт для систем автоматической обработки чувашских текстов, в частности, обеспечивал бы единообразное представление информации, составил бы теоретическую основу морфологической аннотации.

Чувашский язык имеет, что важно для грамматической разметки, развитую систему грамматически однозначных словоизменительных аффиксов (отдельно взятый аффикс выражает один

[1]Публикация подготовлена в рамках поддержанного РГНФ научного проекта № 15-04-00532.

морфологический признак, в нем отсутствуют различные парадигматические классы в парадигме того или иного одного типа; отсутствуют столь значительные чередования в основах, а также закономерная фонетическая обусловленность алломорфов (границы морфем четкие: к основе присоединяются аффиксы с тем или иным значением, а если происходят фонемные изменения на границах морфем, то данные морфонологические изменения связаны с фонологическими законами чувашского языка. При автоматическом анализе аффиксального состава словоформ в рассматриваемом языке грамматические (морфологические) признаки распознаются относительно легко. На первом этапе создания Национального корпуса чувашского языка морфологическая разметка должна содержать, на наш взгляд, информацию о морфологических категориях, явно выраженных аффиксами.

В связи с этим следует отметить, что в условиях существующей чувашской орфографии и действующих принципов освоения заимствований возникает проблема автоматической обработки этих случаев. Необходимо решить, как обрабатывать заимствования, и найти способы описания возможных закономерностей их изменения. Также при автоматической разметке возникают трудности, связанные с полифункциональными и омонимичными аффиксами, с особенностями изменения отдельных категорий слов (в частности, некоторых местоимений и послеложных слов), с частеречной принадлежностью слов в безаффиксальной форме, с числом падежей [7. С.70] и т.д.

В чувашском языке выделяются следующие базовые морфологические классы: имена, глаголы и неизменяемые части речи [5, 6], однако и в них оговаривается, что разграничения между морфологическими классами выражены нечетко, особенно это касается разрядов имен, в частности, слово может квалифицироваться как существительное, прилагательное – все зависит от его синтаксической роли: имя существительное выступает как определение: *йывăç пӳртсем* ‘деревянные дома’; прилагательное может выступать в предложении в роли любого члена: *аслисене итлемелле* ‘старших надо слушаться’.

Зыбкость границ между словоизменительными классами в ходе разработки алгоритмов морфологической разметки чувствуется не только в пределах имени, в частности, если к имени в чувашском

языке присоединяются вербальные показатели (глагол изменяется по лицам), то он принимает в зависимости от синтаксических функций лично-числовые аффиксы – принадлежности (*манăн аллăм* ‘моя рука’, и в составе именного сказуемого маркеры лица (*эпир чăвашсем* ‘мы чуваши’, *-сем* (афф. – 1 л. мн.ч.), в предложении прилаг. в роли актанта принимает именные показатели: *аслисене хисеплеççĕ* ‘старших уважают’, а в случаях выполения атрибутивной функции – функционирует в классе неизменяемых (*чул пӳртсем* ‘каменные дома’).

Лексические формы, относящиеся традиционно к разным морфологическим классам, часто принимают аффиксы одних и тех же морфологических категорий, что для программы является формальным основанием отнести их к одному словоизменительному классу.

В свете сказанного отметим, что модели словоформ верифицировались в монографиях чувашских лингвистов В.И. Сергеева, И.П. Павлова, Н.А. Резюкова, Ю.Н. Исаева и др., которые нужно дополнить инструментами так называемой «грамматики порядков», которые применяются анализе агглютинативных языков [1, 4, 9]. В Грамматике порядков, как удобном инструменте описания агглютинативных языков, морфологии присущи такие требования, как грамматическая однозначность, фиксированная последовательность словообразовательных аффиксов и однократное появление в той или иной словоформе аффикса определенной граммемы.

Для разработки морфологического стандарта в первую очередь, как нам кажется, необходимо описать систему морфологических параметров частей речи. Морфологический анализ тюркских словоформ, как известно, выделяет множество морфем слова и определяет порядок их следования в словоформе, т. е. морфотактические компоненты. С точки зрения разработчика тюркского морфологического анализатора база данных может состоять из лексикона (словарь корневых и аффиксальных морфем), морфотактических и фонологических правил, ибо фонология – это связь между лексической формой слова (уровень глубинного представления слов) и их реализацией (поверхностный уровень).

Обоснованным на первоначальном этапе представляется создание словаря языка, который содержал бы частеречные пометы, не описанные фонологическими правилам чередования основ;

компьютерная модель словоформы; морфотактические правила сочетаемости в пределах словоформы, фонетические правила (выбор алломорфов конкретного аффикса). Словарь языка (основ) должен представить размеченную базу данных, содержащую леммы – слова в начальной форме, ибо морфологическая именная информация в разметка текста включает такие пометы: а) словарная форма лексемы (основа словоформы), б) словоизменительные характеристики (число, падеж существительного); в) справка об исключительной нестабильной морфологической форме, о нарушениях орфографических принципов. Словарь может извлечен из Обратного словаря чувашского языка с использованием технологий СУБД MySQL.

Как и в других тюркских языках, в чувашском, основы не «искажаются», однако в нем проявляются некоторые морфонологические явления, связанные с изменениями основ, в частности, чередование конечных согласных *-у*, *-ӳ* с - ӑв, -ӗв в существительных: çыру – письмо, çырӑва – письму, письмо (дат.-винит. падеж); вӗренӳ – учение, учеба, вӗренӗве – учебе, учебу, учению, учение (дат.-винит.падеж); выпадение гласных и согласных: пурнӑç – пурӑнӑç (жизнь); пыр – пытӑм (пришел).

Введение фонологических и морфотактических правил сыграли бы немаловажную роль в морфологической разметке информацией о грамматических категориях. В частности, учитывая особенности морфотактики в чувашском языке, грамматические параметры именных слот можно подразделить, с одной стороны, на сложные и простые, а с другой стороны, на обязательные и факультативные. Что касается обязательности или факультативности параметра, то обязательный параметр должен быть приписан любой словоформе определенного морфологического класса (, в частности, сущ. обязательно стоят в каком-либо падеже, нет «беспадежных» форм существительных). Факультативный параметр, помимо какого-либо конкретного значения, может также принимать отрицательное значение (отсутствие признака - существительные в чувашском языке могут и не выражать значение принадлежности). В морфотактической схеме рассматриваемого слота участвуют также классы аффиксов, связанных с факультативными параметрами:

N POSS – класс аффиксов поссесивности ЛФ: пурт + ***ӗм (ӑм, у, ӳ , ӗ, и);***ПФ: пуртем;

N INTENSIV - класс аффиксов нтенсивности: ЛФ: пурт + *ах (ех, х);*ПФ: пуртех ;

N PHAZA - класс фазовых аффиксов ЛФ: ир + ***чен (ччен);*** ПФ: ирччен;

N ITERATIVE - класс **итеративных** аффиксов ЛФ: ир + **серен** ПФ: ирсерен;

N LOCATIVE - класс локативных аффиксов ЛФ: хула+н+ ***а + лла (+ е + лле);*** ПФ: хуланалла и др.

Отметим также, что составные классы, связанные со сложными параметрами (как морфологический стандарт используются в разных типах поисковиков для сложной поисковой индексации) принимают активное участие также в морфотактической схеме слота имен, в частности, в слоте «Имя существительное». Рассматриваемые классы, образуемые с помощью послелогов и послеложных слов, имеют структуру (N+Aff Pp или N+Aff PW – здесь N лексемы, имеющие именной тип присоединения аффиксальных морфем, Pp – послелоги, PW – послеложные слова):а) N+Aff ***евĕр*** **(**наподобие, подобно: как будто, словно, как) Pp: лакпа витнĕ ***евĕр*** «словно покрыты лаком», хула çыннисем ***евĕр*** «наподобие городским», пĕр ***евĕр*** хусканусем «одинаковые движения»; б) N+Aff ***валли (***для, на, к) Pp: ĕçлекенсем валли «для трудящихся», кашни ыйту валли унăн ответ хатĕр «на каждый вопрос у него готов ответ» *виçĕ сехет валли кил* «прийти к трем часам»:в) N + Aff ***витĕр (***через, сквозь, в) Pp: *путвал витĕр* «через подвал», *юрпа çил витĕр* «сквозь снег и ветер», *кантăк витĕр пар* «передать в окно» и др.

Схему фонологических правил можно представить следующими компонентами: 1) связь-соответствие между лексическим и поверхностным символами, в частности, соответствие м:н – çын+сем – çын+сен; 2) определяющий данное соответствие операторы (для обозначения их используются математические символы): а) => проявление соответствия только в данном контексте, но не всегда; б) <= проявление соответствия в данном контексте всегда, но не только в нем; в) <=> проявление соответствия в данном контексте всегда и только в нем. Используя соответствующую схему можно описать все чувашские фонологические правила, которые будут отражать все фонологические явления в чувашском языке: законы сингармонизма, гармонию согласных, а также случаи и исключения, возникающие при «осложнении» ЛФ, рас-

смотренных выше смысловыми отношениями (кстати, если внедрить функциональный подход и в фонологические правила, то это послужило бы базой для синтаксического анализатора (СА) для снабжения чувашских текстов детальной не только морфологической, но и предбазой синтаксической информации, когда даже современные существующие СА (в системах Dialing, ЭТАП и др.) эту базу создают с «нуля», имеем ввиду отношения «субъект – действие», «объектные отношения» «темпоральные отношения», «каузальные отношения» и др. (если учесть то обстоятельство, что архитектура тюркских синтаксических анализаторов в корне отличаются от русских в связи с их морфотактикой, автоматной морфологией и их функциональностью, что минимизирует проблемы, связанные со структурой текста (синтаксиса). Уместно отметить и то, что такой подход установил бы также «задел» для актуальной в современной лингвистике (и не только в лингвистке) для систем автоматизации построении онтологии

Модели лингвистической информации, семантико-грамматические аннотации лексем, построенные на основе типологических особенностей чувашского языка можно представить в созданной лексикографической базе инверсионного, грамматического словаря – Обратного словаря чувашского языка [3, 6], ибо практическая значимость словарей такого типа заключается в группировке слов по одинаковому концу: для чувашского языка данный принцип особенно важен, так как аффиксы в нем располагаются справа от корня. Слова в инверсионном словаре в дальнейшем можно сгруппировать по морфологическому признаку (часть речи, наличие или отсутствие того или иного суффикса). В частности, анализ существующих Обратных словарей на практике позволил нам представить многообразие суффиксальных средств имен в чувашском языке, их продуктивность. В обратном словаре имеются массивы слов (более тысячи), которые имеют определенный суффикс. Следует обогатить будущий Грамматический словарь чувашского языка материалами новых словарей чувашского и русского языков, за счет чего расширится круг представленных в них характеристик. В перспективе обратный словарь с программной точки зрения может представить как базу данных и диалоговый интерпретатор запросов к базе данных в разработанной нами ранее системе Java (в этом смысле имеются опережающие нас работы, связанные на только с

чувашским языком, но и в сопоставлении с русским) [2]. Обратный словарь с перечисленными нами характеристиками, т.е. с наиболее полной информацией о грамматической характеристике лексики чувашского языка может быть использован для квантитативного описания по широкому кругу морфологических характеристик, а также значительно расширил бы информационную базу Национального корпуса чувашского языка.

ЛИТЕРАТУРА

1. Володин А.П., Храковский В.С. Типология морфологических классификаций глагола (на материале агглютинативных языков) // Типология грамматических категорий: Мещаниновские чтения. М.: Наука, 1975.

2. Дмитриев А.П., Алексеев М.Б. База данных, применяемых в программе чувашско-русского перевода. Труды Казанской школы по компьютерной и когнитивной лингвистике. Казань, 2010. С. 67–71.

3. Желтов П.В. Сопоставительно-сравнительное исследование морфем чувашского языка с применением формальных методов: диссертация ... кандидата филологических наук. Чебоксары, 2010. 194 с.

4. Ревзин И.И., Юлдашева Г.Д. Грамматика порядков и ее использования // Вопросы языкознания, 1969, N 1, с. 42–56.

5. Резюков Н.А. Сопоставительная грамматика русского и чувашского языков. Чебоксары: Чувашгосиздат, 1959. 327 с.

6. Сергеев В.И. Именные части речи. Чебоксары: Чуваш. ун-т. 1994. 152 с.

7. Сулейманов Д.Ш. К вопросу о числе падежей в татарском языке. Исследования в лингвистике. Казань, 1996. С. 70

8. Унификация систем грамматической разметки в корпусах тюркских языков (семинар UniTurk). Казань, 2014.

9. Gleason H. Introduction to descriptive linguistics. – New York: Holt, Rinehart and Winston, 1955. – P. 503.).

SOME POSSIBILITIES OF SEMANTIC AND ETYMOLOGICAL TAGGING OF CORPORA FOR TURKIC LANGUAGES[1]

Anna Dybo, Alexandra Sheymovich, Sergei Krylov

Institute of linguistics, RAS Moscow, Russia

This paper describes a work on semantic annotating of a corpus of the Khakass language, the design of the inventary of semantic tags. We show examples of tasks which could be resolved using a semantically annotated corpus. The work follows the framework of the RAS corporate project in regards to the development of corpora for languages of the Russian Federation, including Turkic minority languages such as the Khakass.

Работа по семантической разметке корпусов миноритарных тюркских языков в настоящий момент проводится в форме расстановки в статьях электронной хакасско-русской словарной базы, созданной на основе Большого хакасско-русского словаря, семантических тэгов.

Семантическая разметка словарной базы и текстового корпуса значительно расширяет возможности пользователя при создании поисковых запросов и улучшает качество результатов поиска (Кустова, Толдова 2009). Она необходима для решения различных задач на множествах лексем, объединенных в семантические поля, или лексико-семантические классы, по признаку обладания одним или несколькими общими семантическими признаками.

Семантическая информация о лексеме представлена в виде набора семантических помет (тэгов) и записана в поле SEMTAG. Ср. рис. 1.

С учетом необходимости унификации семантической маркировки всех имеющихся национальных корпусов, мы стремимся сделать систему семантических помет для корпуса хакасского языка максимально универсальной, для чего используем существующий опыт и наработки в этой области. В частности, нами были использованы следующие классификации лексики:

[1] Работа ведется на средства гранта РГНФ № 15-04-12030 «*Система автоматического морфологического и синтаксического анализа для корпусов миноритарных тюркских языков России*».

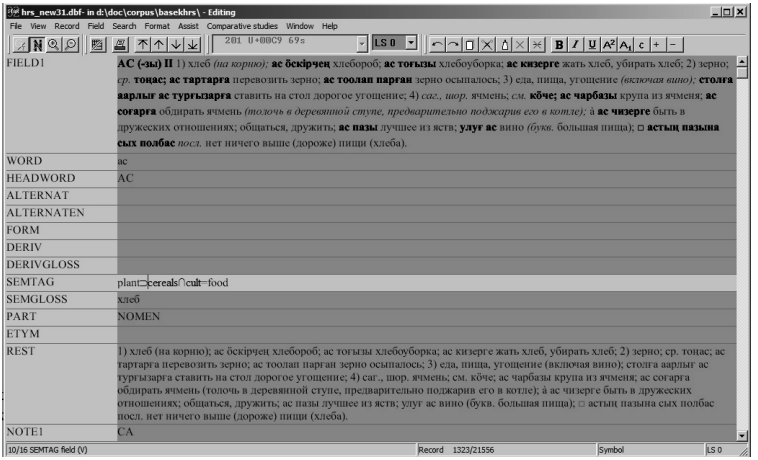


Рис. 1

Тезаурусные, для предметных имен: A.L.E.; СИГТЯ (Лексика); НКРЯ;

Для предикатных имен и глаголов: НКРЯ; частично: Апресян и др. 2007; Апресян 1967.

Общие подходы к организации инвентаря семантических помет

К настоящему моменту разработана предварительная версия инвентаря семантических тэгов; в отличие от помет в НКРЯ, она учитывает не только парадигматические, но и синтагматические характеристики семантики слова. Здесь мы исходим из следующих соображений.

Идея представления языковой семантики через фреймы (или, в более традиционной для российской лингвистики терминологии, предикаты) не является чем-то специфически новым. Предикатное представление лексической семантики было принято как в модели «Смысл – Текст» (ср. у Ю. Д. Апресяна: «В общем случае толкуемой единицей должно быть не отдельно взятое слово, а содержащее его выражение вида *XPY*, где *P* – толкуемое слово, а *X* и *Y* – переменные, сообщающие данному выражению форму предложения или словосочетания» – Апресян 1968, с. 97; еще раньше – в работах А. К. Жолковского, Н. Н. Леонтьевой и Ю. С. Мартемьянова,

с 1960 г.), так и в различных направлениях генеративистики (несколько позже – примерно с 1963 г., поскольку генеративная грамматика далеко не сразу признала необходимость работы с семантическим уровнем языка). Пожалуй, наиболее подробная разработка «фреймовой» семантики сейчас принадлежит Ч. Филлмору и его сотрудникам, см. Berkeley Framenet Project, http://framenet.icsi.berkeley.edu/ framenet (сам Ч. Филлмор вел исследования в этом направлении с 1982 г.)[1].

Вкратце современные лингвистические представления о «фреймовой» семантике можно очертить следующим образом. Слово естественного языка – знаковая единица, то есть двусторонняя единица, имеющая форму и значение. Отдельные значения многозначного слова рассматриваются при семантическом анализе как составные части отдельных единиц. По крайней мере для значительной части слов – лексических единиц (ниже мы вернемся к вопросу – для какой части) следует считать, что каждое из них описывает стандартизованное представление о некотором типе реальных ситуаций. Типизованная ситуация («фрейм») включает «участников ситуации», они же «семантические роли», связанные элементарными отношениями. Семантические роли делятся на «ядерные» («обязательные») – сущностно важные для ситуации, такие участники, без которых сама ситуация не может иметь места, и которые характеризуют в данном наборе именно этот тип ситуаций; и «неядерные» – не характеризующие специфически данный тип ситуаций (хотя, возможно, и необходимо присутствующие в нем). Пример – ситуация торговли; она включает в качестве ядерных роли покупателя, продавца, акт передачи (сам по себе может быть описан как фрейм с более элементарными участниками), предмет торговли («ресурс» в терминологии В.Ю. Розенцвейга 1964), деньги (или иной кодифицированный эквивалент обмена). «Неядерные» роли здесь – время, когда совершается сделка, место, где она совершается (роли, характеризующие любую ситуацию типа события или процесса), цель сделки (в принципе, может отсутствовать); о последнем типе «неядерных» ролей иногда говорят как о «внешних» ролях. Типы ситуаций могут быть расклассифицированы внутри тезауруса, соответственно, вступая

[1] Более подробно о этом классе семантических теорий см. в: Лингвистика конструкций (изд. Е.В.Рахилина), М., 2010, с. 18-75.

в отношения включения, пересечения и подчинения между собой. Торговля – подтип ситуации обмена, в ситуации обмена не специфицирован как кодифицированный эквивалент второй ресурс; купля и продажа – подтипы ситуации торговли, в зависимости от того, покупатель или продавец рассматривается как главный герой ситуации; спекуляция – еще один подтип ситуации торговли, где участник «цель» включен как внутренний – торговля с целью получения неоправданно высокой прибыли (для нас неважно, что реально соответствующая ситуация, возможно, ничем не отличается от просто торговли – здесь речь идет об обычном значении слова). Предполагается, что словарь каждого языка должен представлять собой базу данных с «фреймовыми» толкованиями для слов, включающую информацию о синтаксической интерпретации «семантических ролей» и (если требует грамматика языка) стандартном морфологическом выражении этой синтаксической интерпретации – то есть, возвращаясь к принятой у нас терминологии, информацию о модели управления слова[1]; кроме того, для каждого фрейма имеются ссылки на все слова, фреймы которых совпадают с данным или находятся с ним еще в каких-либо отношениях. Имея такой словарь, автоматическая система может производить «аннотирование» предложений любого текста, то есть

[1] Вообще-то, различие между традиционным лексикографическим толкованием и толкованием в предикатной форме носит скорее технический характер; в общем случае из традиционного толкования можно получить «фреймовое» автоматически; для русского языка для этого следует перевести инфинитив правой части толкования в личную форму и приписать справа и слева от нее переменные X и У к ближайшим двум зависимым существительным, далее Z и т.п.; эксперимент был проделан С.А.Крыловым над толкованиями словаря Ожегова с помощью системы автоматического морфологического анализа, встроенной в СУБД Starling (система управления этимологическими базами данных, разработанная С.А.Старостиным, http://starling.rinet.ru) и дал вполне удовлетворительные результаты. Другое дело, что традиционная лексикография не требует в эксплицитной форме подробного описания модели управления слова, почему в словарях языков с менее развитой лексикографической традицией, чем русская, французская, немецкая или английская (а лучше всего – латинская и древнегреческая) зачастую моделью управления слова вообще пренебрегают, а «предикатная» лексикография специально заостряет внимание лексикографа на этой проблеме, эксплицитно рассматривая описание возможных типов управляемых единиц как основную часть толкования управляющего слова.

относительно каждого выбранного слова («мишени») размечать в тексте обозначения «участников» соответствующего фрейма, что и моделирует процесс понимания предложения и, в дальней перспективе, текста. Соответственно, при наличии таких баз для нескольких языков можно смоделировать автоматический перевод с языка на язык.

Из изложенного видно, что легко поддаются «фреймовому» толкованию языковые предикаты, то есть слова, обозначающие действие, состояние или процесс. Со словами других типов возникают сложности, разрешение которых производится различными способами, ad hoc. Наибольшую трудность для осмысленной фреймовой интерпретации представляют так называемые предметные имена, так сказать, термы par excellence. Надо сказать, что их толкование вызывает сложности практически во всех семантических теориях (ср. традиционный для лексикографии вопрос о том, насколько словарное толкование должно включать энциклопедическую информацию)[1].

Общий теоретически мыслимый подход к толкованию предметных имен таков: во всяком случае, для имен артефактов и иных предметов, регулярно участвующих в человеческой деятельности, явно можно использовать классы типа «материал», «инструмент», «помещение», «сосуд» и под. – каждый из таких классов легко интерпретируется как фрейм (т.е. предикатно, с более элементарными участниками); понятно, что отнесение предметного имени к одному из таких классов отражает наиболее типичное его использование в человеческой деятельности. По-видимому (как показывают, во всяком случае, историко-семантические исследования), в реальных языках подобным же образом устроена и семантика ряда наименований природных объектов – например, названий расте-

[1] Мы не будем здесь специально уделять внимания частной проблеме в семантике естественного языка, а именно, связи различных типов значений многозначного предметного имени с его типичной синтаксической функцией, на которую впервые обратил внимание В.В.Виноградов в статье «О типах лексических значений» (1953 г.; у имен типа «свинья» основное значение – конкретное, производное – метафорическое, и второе встречается практически исключительно в синтаксической функции предиката). В дальнейшем Н.Д. Арутюнова дала уточненную классификацию типов «таксономического» и «характеризующего» значений; в системе фреймовых толкований у Филлмора «характеризующие» значения также обрабатываются особым образом.

ний и животных, частей тела («конопля – травянистое растение, используемое для изготовления веревок»; именно такая семантика приводит к тому, что в истории языков название конопли часто переходит в название крапивы и наоборот). Неудобство такого подхода к фреймовым толкованиям, очевидно, то, что предложения, описывающие ситуацию нетипичного использования объекта и потому не поддающиеся простому автоматическому «аннотированию», будут встречаться значительно чаще и поддаваться дополнительной типизации значительно меньше, чем предложения, описывающие ситуации отступления от словарных фреймов глагольного типа (вроде случаев с разнотипными опущениями участников или наложениями нескольких разных фреймов), просто в силу того, что глагольные предикаты – в сущности, продукт именно человеческой интерпретации ситуаций, деятельности человеческого разума, а множество предметных имен – понятийная сетка, с трудом налезающая на множество объектов действительности, не являющееся само по себе чем-то специально приспособленным для человеческой деятельности, и потому значительно более разнообразное.

Соответственно этому, техническое решение, которое принимает команда Филлмора (и, по-видимому, наиболее естественное) – следующее. Создание фреймовой базы данных производится не «вообще», а для «аннотирования» текстов из определенного корпуса. При толковании предметных имен сначала выявляют наиболее частотные в данном корпусе предикаты, с которыми употребляется данное предметное имя, соответственно, выясняется, в каких фреймах значения этих имен типичным образом заполняют «семантические роли». Соответственно, для имени строится фрейм с данным типом фреймов в качестве «участника», а также, по возможности, с еще одним или несколькими «участниками», представляющими типизованную «качественную» характеристику (вроде «материала», из которого предмет изготовляется, или «цели», для которой он предназначен)[1].

[1] Собственно говоря, сам Филлмор не решается говорить о предикатных (фреймовых) толкованиях предметных слов, утверждая, что при аннотировании текста отдельным образом производится аннотация для мишени-носителя фрейма и мишени-»заполнителя слота», но в действительности толкования «заполнителей слотов», устроенные так, как было описано выше, легко записать как предикатные.

Такой подход, в общем, соответствует теоретико-семантическому представлению о «возможных мирах», в рамках которых только и может быть правильно проинтерпретирован смысл той или иной лингвистической единицы. Соответственно, тематически ограниченный корпус текстов может рассматриваться как реализация своего рода «платоновской идеи» некоторого возможного мира, внутри которого происходит понимание текста и слова текста получают толкование.

Разметка хакасского словаря (и впоследствии автоматических словарей других миноритарных тюркских языков) предназначена для предварительной работы с корпусом текстов, разнообразных по тематике; статистическая обработка текстов с выяснением сочетаемости слов планируется впоследствии, но видится как один из промежуточных результатов, а не исходный пункт работы. Поэтому пока мы даем синтагматическую информацию в тэгах только для глаголов, а также очевидных операторов, в частности, имен действия, состояния и процесса. В дальнейшем планируется расширить этот подход к пометам; пока же мы ориентировались в целом на подход, декларированный в статье (Апресян и др. 2005, 193—214). Согласно ему, инвентарь семантических тэгов (или дескрипторов), своеобразный семантический метаязык, должен обеспечивать содержательное и адекватное описание лексики языка, предметной и предикатной[1], а также, в совокупности с морфологической и синтаксической разметкой, давать исследователю достаточную информацию о закономерностях поведения элементов всех лексико-семантических классов в текстах на естественном языке. Предметные тэги членят словарь языка с наивно-энциклопедической точки зрения, отражая некоторым образом систему представлений носителя языка об окружающем мире. Поэтому, к слову сказать, так удобно было использовать для разработки сем. инвентаря том СИГТЯ «Лексика», организованный по принципу тезауруса и содержащий разделы, соответствующие

[1] Предметная лексика – названия живых существ, растений, гор, рек, овощей, фруктов и т.п., предикатная лексика – любые другие валентные лексемы, главным образом обозначающие действие, состояние или процесс, дескрипторами для которых будут 'действие', 'деятельность', 'занятие', 'воздействие', 'свойство', 'интерпретация' и такие их подклассы как 'начало' и 'прекращение', 'каузация' и 'ликвидация'. Апресян и др. 2005

важнейшим сторонам природы, жизни и хозяйства носителей древних диалектов пратюркского языка.

Об инвентаре семантических помет

Наша система семантических помет выглядит иерархически: тэги вступают в отношения включения, пересечения и объединения, а также отношения «Аргумент – Функция» между собой. Для символизации отношений между признаками мы воспользовались стандартной теоретико-множественной и логической символикой:

⊃ – включение, помета, следующая за этим знаком, конкретизирует предыдущую (*hum ⊃ persn , hum ⊃ prof*); в словарной статье конкретизируемый признак может опускаться.

∩ – пересечение признаков (*plant∩cult*).

& «и» – конъюнкция признаков (*period&quant(time)*)

∨ "или" – дизъюнкция признаков (*hum∨ animal*).

Аргументы признаков-операторов ставятся в скобках после операторов (*part(body), part(plant), part(constr)*); скобки же размечают соподчинение связок.

| – отделяет сведения о валентностях предиката;

: двоеточие – отделяет заполнения валентностей (|*Ag:hum, Pat:plant*).

= – разделяет пометы, принадлежащие разным значениям многозначного слова (*plant∩cult=food*).

Соответственно, слово может попадать в несколько семантических классов (т.е. одно слово м.б. снабжено несколькими наборами помет – см рис. выше (так же устроены пометы в НКРЯ)). Слова, принадлежащие к разным лексико-грамматическим классам, могут оказаться в одном семантическом классе:

HEADWORD **ААЛ I** село, селение; населённый пункт; аал; улус // сельский;
SEMTAG settl=aggr(hum)

HEADWORD **АҒАС (-зы)** 1. 1) дерево; ... 2) древесина
SEMTAG plant=stuff

HEADWORD **АС (-зы) II** 1) хлеб (на корню); ... 2) еда, пища,
SEMTAG plant∩cult=food

HEADWORD **ЧИДІЛЕРГЕ /чиділ-/** кашлять;
SEMTAG physiol⊃disease|*Pat:hum*

HEADWORD **ЧИДІЛ** кашель;
SEMTAG physiol⊃disease

Ниже следует фрагмент инвентаря семантических помет для хакасской словарной базы, составленный с опорой на вышеприведенные источники; при некоторых пунктах даются примеры сложных обозначений:

а) оппозиция собственные имена *onym* vs предметные имена *concr* vs имена предикатов *abstr*

Onym ⊃ (persn; patrn; famn; topon)
имена *hum* ∩ *persn*
отчества *hum* ∩ *patrn*
фамилии *hum* ∩ *famn*
топонимы landsc ∩ *topon*

б) тезаурусная классификация

I. Природные явления и объекты *nature*
1. Небо и небесные тела *astr*
2. Атмосферные явления; ветер, погода *weather*
3. Вода; моря, реки *landsc* ⊃ *water*
4. Земля, недра, почвы *soil*
5. Вещества и материалы *stuff*
4. Ландшафт *space* ⊃ *landsc*
6. Времена года, сезоны *season&period (time)*
7. Растительный мир
7.1. Дикорастущие деревья, кустарники, травы *plant*
[цветущие растения *plant*∩*flover*]
[Культурные растения *plant*∩*cult*]
[Плодовые и ягодные деревья и кустарники *plant*∩*fruit*]
7.2. Земледельческие культуры, злаки *cereals*∩*cult*
7.3. Огородные культуры *veg*∩*cult*
7.4. Плоды *fruit*
7.5. Цветы *flower*
7.6. Деревья *tree*
7.7. Кусты *bush*
7.8. Трава *grass*

8. Животный мир *animal*
8.1. Дикие животные *animal* ⊃ *beast*
[Домашние животные *animal* ⊃ *beast*∩*dom*]
8.2. Птицы *animal* ⊃ *bird*
[Домашняя птица *animal* ⊃ *bird*∩*dom*]
8.3. Рыбы *animal* ⊃ *fish*
8.4. Земноводные и пресмыкающиеся *animal* ⊃ *reptile*
8.5. Насекомые *animal* ⊃ *insect*

II. сверхъестественное *supernat* /мифологическое *myth*
[джинны и пери *hum*∩*supernat*; драконы *animal*∩*supernat*]

III. Человек
1. Наименования лиц *hum* в том числе:
1.1. имя представителя профессии *hum* ⊃ *prof*
1.2. этнонимы *hum* ⊃ *ethn*
2. Тело *body(hum)*, *body(animal), body (hum* ∨ *animal)*
[части тела *part(body)*
Органы *part* ⊃ *organ(body)*]
3. Физиология человека и животных *physiol*
4. болезни *disease*

IV. Виды и области человеческой деятельности (*activity*)
1. Сельское хозяйство: земледелие и работа на земле *cult*
1.1. Выпечка хлеба *bread*
2. Животноводство *dom*
2.1. Молочное производство *milk*
2.2. Мясное производство *meet*
2.3. Пастбище *pastur*
3. Охота, рыболовство, обработка рыбы, мяса *hunt*
3.1. Собирательство *gather*
4. Еда и напитки, курение *food*
5. Материальная культура *artefact*
5.1. Гончарное дело *ceramic*
5.2. Деревообработка *woodwork*
5.3. Металлы, обработка металлов, кузнечное дело *metal*
5.4. Текстильное и кожевенное производство *textil*
5.4.1. Одежда, обувь, головные уборы *cloth*
5.5. Изготовление войлока *felt*
5.6. Поселение, жилище *settl*
5.6.1. Строительство: здания и сооружения *constr*

5.6.2. Дом, детали устройства *house*
5.7. Инструменты и приспособления *tool*
5.7.1. механизмы и приборы *device*
5.7.3. музыкальные инструменты *mus*
5.7.4. мебель *furn*
5.7.5. Кухонная утварь, посуда *dish*
6. Огонь, добывание огня *fire*
[Обогрев *fire*∩ household;
Освещение light∩household]
7. Работы по дому *household*
8. Дороги и транспорт, транспортные средства *transp*
9. Общественная деятельность *social*
9.1. государство *state*
9.2 Война, армия *war*
9.2.1. оружие *weapon*
9.3. Организации *org* (больница, школа, музей)
9.4. Торговля *trade*
9.5. финансы *finance*
9.6. Религия *relig*
9.7. Наука *science*
9.8. Искусство *art*
9.9. Образование *educ*
9.10. мероприятия *event* (*аукцион, вернисаж, вечеринка, выборы, именины, заседание, культпоход*)
9.11. игры *game*
9.12. спорт *sport*
9.13. праздники *holid*
V Человеческое поведение *behav* (*куролесить, привередничать*)
1. взаимодействие и взаимоотношение *inter(hum)* (*взаимопомощь, вражда, схватка, драка*)
2. Ментальная сфера
2.1. Мышление *ment*
2.2. память *ment* ⊃ *mem*
3. Психическая сфера *psych*
3.1.эмоция *psych* ⊃ *emot*
3.2. воля *psych* ⊃ *volit*
3.3. восприятие *perc*

11. Язык и коммуникация *lang*
11.1. речь *speech*
11.2. тексты *text*

VI. Общественные отношения *rel*
1. Родство *kin (термин родства hum∩kin)*
2. Брак *matrim*
3. Свойство́ *affinity (термин свойства hum∩affinity)*

VII Меронимия
1. части в том числе: *part()* (напр., *part(hum), part(constr)*)
2. кванты и порции *qtm()* (напр., *оборот, прыжок, кивок*; в т.ч. вещества, времени: *qtm(time)*)
3. множества *set()* (*система, выборка, алгоритм*)
4. совокупности объектов *aggr()* (*толпа*)
5. имена классов *class()* (*ворьё, пролетариат*)

VIII. Топология
1. вместилища *contain*
2. горизонтальные поверхности *horiz*
3. Верх *super*
4. Низ *sub*
5. право *dextr*
6. лево *sinstr*
7. Внутри *in*
8. Снаружи *extr*

IX. Оценка *ev* в том числе:
1. положительная оценка *ev ⊃ posit*
2. отрицательная оценка *ev ⊃ neg*

X. Движение *move*
1. изменение положения тела, части тела *move (body)*
2. помещение объекта *put*

XI. Действие *action*
1. физическое воздействие *impact*
2. создание физического объекта *impact ⊃ creat)*
3. уничтожение физического объекта *impact ⊃ destr*
4. взаимодействие *inter*
5. применение *Oper*

XII. Изменение состояния или признака *changest*
1. состояние *state*
2. процесс *process*

3. функционирование *Func*

XIII. бытийная сфера *be*

1. существование *exist*
2. начало существования *be* ⊃ *appear*
3. прекращение существования *be* ⊃ *disapp*

XIV. Каузативность и подр. *Caus*

Что именно «это всё»? Полуслужебные глаголы, служащие предикативизаторами (операторами образования предикаций из) абстрактных существительных?

XV. Пространство, место и пространственные отношения

1. местонахождение *place*
2. положение тела в пространстве *space*
3. контакт и опора *contact*

XVI. Обладание *posess*

XVII. Качества *qual*

[*qual(hum)* человеческие качества, *qual(tool)* качества инструментов,

qual(body) качество тела – сильный, толстый и т.д.]

XVIII. Физические свойства *phys*

1. звук *sound* [звукоподражания *sound*∩*speech*; “слушать/слышать” *perc(sound)*]
2. цвет *colour* [цветообозначения *colour*∩*speech*]
3. свет *light*
4. запах *smell*
5. вкус *taste*
6. температура *temper*
7. вес *weight*
8. форма *form*
9. размер *size*
10. Расстояние *dist*
11. Скорость *speed (move)*

XIX. Время (*time*) в том числе:

1. период *period*
2. момент *moment*
3. день недели *week*
4. месяц *month*
5. возраст *age*
6. Длительность *dur(period)*

XX. Исчисление *numer*
1. Величина
в том числе:
большая *max*
малая *min*
абсолютная *abs*
2. Количество *quant*
3. Единица измерения *unit*
4. Параметр *param* [*высота, грузоподъемность*]
XXI. Модальность *mod* [надобность, возможность, нужно]
XXII. ЧИСЛИТЕЛЬНЫЕ *num*
1. количественные *num* ⊃ *card*
2. порядковые *num* ⊃ *ord*
3. собирательные *num* ⊃ *collect*
XXIII. МЕСТОИМЕНИЯ *pron*
1. личные *pron* ⊃ *pers* (*я, он*)
2. возвратные *pron* ⊃ *ref* (*себя*)
3. притяжательные *pron* ⊃ *poss* (*мой, его, свой*)
4. вопросительные/относительные *pron* ⊃ *rel* (*кто, который, когда*)
5. указательные *pron* ⊃ *dem* (*этот, такой*)
6. неопределенные *pron* ⊃ *indet* (*некоторый, некогда*)
7. отрицательные *pron* ⊃ *neg* (*никакой, ничей*)
8. кванторные (определительные) *pron* ⊃ *spec* (*всякий, каждый, любой*)

Отношения семантической и словообразовательной разметок

Отдельно следует сказать о существенной для нашей базы особенности расстановки частеречных и семантических помет и различиях в их трактовке. Как известно, существуют некоторые отличия в способе выделения частей речи в тюркских языках различными и грамматиками, скажем так, тюркски и европейски ориентированными. Напомним, что одно и то же слово в тюркских языках может трактоваться как существительное, или прилагательное, или наречие в зависимости от синтаксической функции, выполняемой им в предложении (*тас тура* ‘каменный дом’ *кічіглер ойнапчалар* ‘маленькие (малыши) играют’). Поэтому в нашей мо-

дели нет разных систем тэгов для традиционно выделяемых частей речи, как, например, в НКРЯ, разработчики которого пишут: «Лексико-семантическая информация имеет различную структуру для разных частей речи» (http://www.ruscorpora.ru/corpora-sem.html). В НКРЯ свою структуру помет имеют собственные, предметные и непредметные существительные, прилагательные, наречия, глаголы. Причем в систему семантических помет включена и грамматическая и словообразовательная информация: каузация и служебный статус у глаголов, диминутивы, аугментативы, сингулятивы и т.п. у имен.

В электронной хакасско-русской словарной базе для информации такого рода предназначено поле DERIVGLOSS, где представлено деление основы на словообразовательные форманты и даны стандартизованные граммемные имена этих формантов, в основном из инвентаря лексических функций Модели «Смысл <=> Текст».

HEADWORD СЫСХАРАРҒА /сысхар-/ 1) растапливать, вытапливать (жир); сосха чаа сысхарарға растопить свиное сало; 2) процедить, отделить вытопки *(напр.,* масла, сала).

DERIVGLOSS *просачиваться&плавиться=Caus-[1]

SEMTAG Caus(process(food ∩ household))

Отметим, что такие операторы как каузация (caus) используются и как имя словообразовательного типа, и как семантический тэг, т.к., с одной стороны, основы, содержащие словообразовательный каузативный аффикс, не всегда являются каузативами семантически:

HEADWORD аарладарға быть уваженным

DERIVGLOSS тяжелый=Oper=Caus-

SEMTAG Pass(inter∩posit)

(семантическое развитие, очевидно, через “вызывать уважение”)

- а, с другой стороны, семантическая каузативность не всегда формально выражена в основе:

---

[1] Относительно редкий каузатив от незафиксированного в хакасском как непроизводного пратюркского глагола **siř-* > **siz-* 'сочиться; таять, плавиться' ЭСТЯ 2003, 394.

HEADWORD ОДЫНАРҒА /одын-/ «топить»
DERIVGLOSS *Inch(огонь)=Refl-[1]
SEMTAG Caus(Func(Fire∩Household))

HEADWORD	**ХЫРАРҒА /хыр-/ I** отскабливать что-л.
DERIVGLOSS	-
SEMTAG	Caus(disapp)
HEADWORD	**ХЫРАРҒА /хыр-/ II** уничтожать, истреблять
DERIVGLOSS	-
SEMTAG	Caus(disapp)

Что касается таких традиционно выделяемых в грамматиках лексико-грамматических классов (и соответствующих им частей речи) как «местоимение» и «числительное», то мы предпочитаем не вычленять их из лексико-грамматического класса «имя». На наш взгляд эти типы имен-шифтеров (ср. Якобсон 1972, 95–99) не являются отдельными грамматическими классами, поскольку набор их грамматических категорий совпадает с именными. Как и другие имена, местоимения и числительные могут употребляться в функциях актантов, атрибутов и предикатов.

ікі *алтам*	**два** шага;
ікі *ал салғам*	[я] получил **двойку**
ікі *хоныхтығ*	**дважды** женатый
ӱс *аймах* ***ӧзім***	**три** вида растений

Пастап ***ол миннең*** *хорыххан, ам пызо****ҷ****ах осхас чабас*
Сначала **он меня** боялся, а теперь как теленок смирный

Мына, ***оларның*** *пірсі Алексей Кружков*
Вот (этот) **один из них** Алексей Кружков

---

[1] Отыменной глагол **ōt-и-* образован с помощью суф. *-и-*, слабо продуктивного в современных тюркских языках, но довольно обычного для древнетюркского, со значением становления признака (т.е. «загораться») – см. Erdal 1991, 474-479. Не зафиксирован в тюркских языках; но имеются производные от него с каузативным суф. (алт. *одыр-* 'разводить огонь') и с рефлексивным (хак. *одун-* 'топить'; др.-уйг. *otun-* 'растапливать' (?) в сочетании *otungu otuŋïn* 'дрова (Acc) на растопку' Maitr., см. Erdal 1991, 609); а также общетюркское производное имя на *–ŋ* (Erdal 1991, 337-338) *otuŋ* 'дрова' (производность на *–duŋ* с фонетич. упрощением, предлагаемая М.Эрдалом на с. 156, маловероятна).

Ол *Тойоңзар, пысхан нымырт осхас, хара харахтарынаң хази көрібіскен,* ***олох*** *көрізін Арина Петровназар тастаан*

Она пронзительно посмотрела на Тойона [своими] как спелая черемуха черными глазами, **этим же** взглядом окинула Арину Петровну

пу *турада ағылахсынчам*
в **этом** доме мне просторно

Пулар түрче хаңаада, алаң ас парып одырчатханнар
Они (эти), растерявшись, недолго сидели в телеге

Эти группы имен представляют собой специфические семантические классы. Поэтому пометы «местоимение» и «числительное», включая их лексико-грамматические разряды (*pron:pers, num:card*), в хакасской словарной базе относятся к полю SEMTAG:

HEADWORD **ОЛ** мест. он, она, оно; тот, та, то
SEMTAG *pron* ⊃ *pers=pron* ⊃ *dem*
HEADWORD **СИГІС** восемь
SEMTAG *num* ⊃ *ord*

Некоторые примеры использования семантических помет

Самый распространенный тип запроса к семантически размеченной словарной базе – это нахождение по данному семантическому. признаку слова или группы слов. Пользователь получает множество лексем, характеризующееся той или иной степенью близости значений (синонимический ряд, гипероним с его гипонимами, антонимы, конверсивы, семантическое поле в версии Ю. Н. Караулова, набор семантических функций от данного слова в версии Мельчука–Жолковского–Апресяна или лексико-семантическую группу в версии Э. В. Кузнецовой).

Второй тип запросов служит решению более сложных научно-исследовательских задач. Это, например, определение лексико-грамматической сочетаемости слов, снятие семантической неоднозначности многозначных слов и т.п.

1. Задача полуавтоматического снятия семантической неоднозначности глаголов на основе сем. характеристик актантов

Задачи снятия семантической неоднозначности глаголов исходя из семантической. разметки контекста, а именно с помощью

данных о глагольных моделях управления, в которых актантам глагола проставлены семантические. тэги, т.е. с помощью т.н. «семантических фильтров» (Кустова Г. И., Толдова С. Ю. НКРЯ: семантические фильтры для разрешения многозначности глаголов // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009, 258—276).

Примеры глаголов-омонимов:

В настоящий момент лексические омонимы и просто многозначные слова в словарной базе различаются римскими и арабскими цифрами в поле FIELD 1. Теперь мы собираемся различать их еще и с помощью семантических помет, как самих глаголов (в словаре), так и их актантов (в размеченных текстах корпуса).

КИЗЕРГЕ /кис-/ I 1) резать, разрезать, срезать, отрезать что-л.; іпек кизерге резать хлеб; пычахнаӈ кизерге резать ножом;
кис- *impact* | *Ag: Pers; Pat: Concr; Instr: Tool*
іпек *food*
пычах *instr*

КИЗЕРГЕ /кис-/ II 1) надевать что-л.; ӧдік кизерге надевать обувь, обуваться; 2) носить что-л.: тон кизерге а) надевать пальто;
кис- *put* | *Ag: Pers; Pat: Cloth*
тон *cloth*
ӧдік *cloth*

КИЗЕРГЕ /кис-/ III переходить что-л.; переправляться через что-л.; суғ кизерге переправляться через реку; чол кизерге переходить через дорогу; чазағ кизерге переходить вброд.
кис- *move* | *Fact: Pers&Anim; Loc: Space*
суғ *landsc* ⊃ *water*
чол *space*∩*transp*

Соответственно, автоматический парсер дает форме глагола с основой *кис-* три варианта разбора, отличающихся номером и переводом глагола, взятыми из словаря. Можно построить фильтр, который будет проверять соответствие семантических помет актантов в словарной статье глагола семантическим пометам имен, встретившихся в том же предложении, что и глагол, в тексте. Как мы видим, в примерах предложений, приведенных в словаре при каждом глаголе, такое соответствие соблюдается. Если при каком-

то из вариантов анализа совпадения нет, такой вариант анализа отбрасывается. Т.е., может быть выведено правило вида: «Если актант глагола *кис-* выражен именем, принадлежащим к сем. классу “одежда”, то в данном контексте значение этого глагола – ‘надевать’». В дальнейшем правило может быть обобщено для целого класса однотипных глаголов; выделение таких классов для наших корпусов представляется делом будущего.

2. Задача полуавтоматического снятия семантической неоднозначности многозначных имен на основе их лексической сочетаемости

Разрешение омонимии в группе имен «час I “слеза” [*physiol∩water*]; II возраст, год, лета [*quant(time)*]; III 1) час // часовой [*period&quant(time)*]; 2) часы [*device*]; IV 1) молодой, зелёный [*age; color(plant)*]; 2) свежий [*qual(food)*]; V сырой [*qual(concr)*]; VI весна [*season*]» не осуществимо автоматически для всех значений многозначного слова (при том, что слово III пополнило эту группу тюркских омонимов, будучи заимствовано из русского – но не обладая формальными признаками русизмов).

По признаку лексической сочетаемости с числительными лучше всего определяются значения ‘год’ и ‘час’ (**апсахха тоғызон час** старику девяносто лет; **пала пис час толдырды** ребёнку исполнилось пять лет), также они встречается в контексте слов с сем. тэгом *time* (**иртен читі часта** утром в семь часов) – зато различие между ними практически невыявимо; значение ‘часы’ (единственное среди них с пометой, входящей в класс Concr) выявляется, например, в контексте слов с пометой *stuff* (**алтын час** золотые часы). Значение ‘зеленый’ можно присвоить слову *час* в контексте слов с сем. тэгом *plant, plant:part* (**час от** молодая трава; **час от чіли пӳктелче, час хамыс чіли мондылча** *фольк.* как зелёная поросль изгибается, как молодой камыш качается *(о пластике движения молодой девушки)*), однако значительно сложнее вычленить значения *час* ‘свежий’, т.к. это может быть и пища (**час халас** свежий хлеб), и ребенок (**час пала** новорождённый младенец) – слова, принадлежащие к различным семантическим группам; такая же ситуация и со значением ‘влажный’ – древесина, шкура, облака (**час ағас** сырая древесина *(как строительный материал);* **час теер** сырая шкура *(только что снятая с животного);* **час пулуттар** тяжёлые, кучевые облака *(дождевые или снеговые).*)

3. ***Отражение числа в глаголе при согласовании подлежащего и сказуемого****.*

В хакасском языке согласование подлежащего и сказуемого в числе зависит от того, чем они выражены. Если подлежащее в единственном числе, сказуемое всегда с ним согласовано (ед. число не маркировано). Если подлежащее во множественном числе, то его согласование имеет некоторые особенности [ГХЯ 1975 303-304]. Согласование по числу обязательно лишь в том случае, если подлежащее выражено местоимениями 1 и 2 л. мн.ч. В большинстве остальных случаев, когда подлежащее выражено 3-м лицом, согласование его со сказуемым происходит факультативно: *Олар тоғынча / тоғынчалар* ‘Они работают’; *Піс (2 л. мн.ч.) часкалығбыс, я?* ‘Мы счастливые, да?’ На последнем примере можно наблюдать забавный случай автоматического снятия лексической омонимии: слову *піс* морфологический анализатор автоматом припишет 2 перевода: личное местоимение *мы* и существительное *шило*, а по признаку *Pers2 Pl* сказуемого можно смело выбрать из двух омонимов местоимение *мы*.

4. Использование семантической разметки отедельных словарей при составлении этимологической базы данных

В процессе семантической разметки хакасской словарной базы стала очевидна необходимость разграничения двух больших, хорошо различимых пластов лексики – собственно хакасской (шире – тюркской) и пришедшей в хакасский через посредство русского языка. Адаптируя уже существующие в различных корпусах инвентари сем. помет к своим нуждам, мы пришли к выводу, что есть необходимость детализировать систему помет для предметных имен, обозначающих, например, растения, животных, виды человеческой деятельности, инструменты, предметы домашнего обихода, профессии и т.п. – т.е. лексику, описывающую традиционные сферы жизни и деятельности носителей хакасского языка (области, в которых он еще не так значительно сдал позиции русскому) Для этих целей в словарную базу было введено еще одно новое поле ETYM, предназначенное для этимологических помет. При первом прохождении словаря на предмет проставления семантических тэгов в этом поле параллельно проставляются пометы rus – только для русских заимствований и слов, пришедших в хакасский через русское посредство (для тюркских по

происхождению слов впоследствии планируется сделать ссылки на тюркскую этим. базу). Инструментарий системы Starling позволяет для значительной части случаев проставить помету автоматически (это, например, все основы, начинающиеся на звонкую букву, и под.). Наложение фильтра по признаку ETYM = rus даст нам следующие рабочие возможности: отделение части словаря, для которой можно позаимствовать сем. тэги из словарей русского языка и НКРЯ, от другой, бóльшей части, состоящей из тюркской лексики; возможность более наглядно оценить степень освоения русских заимствований хакасским языком не только в семантическом плане, но и в грамматическом (особенности слово- и формообразования; моделей управления). В дальнейшем этимологические пометы должны быть проставлены по всей базе, это даст возможность использовать ее в работе по параллельному проекту Полной этимологической базы по тюркским языкам, и в перспективе планируется автоматически связать между собой все наши тюркские базы. Семантическая разметка в сочетании с наложением этимологических фильтров может помочь автоматическому выбору между омонимами при установлении взаимосвязей между этимологической базой и словарями конкретных языков. Для этой работы программа должна иметь семантическую информацию для разрешения омонимии (или квазиомонимии). Семантическая разметка окажет значительную помощь при определении семантической близости родственных слов из словарных баз различных тюркских языков. Так, можно будет сравнить, насколько хорошо сохраняются др-тюрк. значения в современных тюркских языках и каково соотношение между ними.

Примеры

Хакасская словарная база

HEADWORD **AT** I лошадь, конь
SEMTAG *animal*∩*dom*

HEADWORD **AT** II имя, кличка
SEMTAG *speech(onym)*

HEADWORD **АТАРҒА /ат-/** I 1) стрелять
SEMTAG *impact*|Instr:weapon

HEADWORD AC I ласка
SEMTAG *animal*

HEADWORD AC II хлеб
SEMTAG *plant∩cult=food*

HEADWORD AC III голод
SEMTAG perc(physiol∩neg)

HEADWORD AC IV мало
SEMTAG min

Подбаза “Древнетюркский” из Этимологической базы по тюркским языкам:

AT, ĀT I 1. имя: ädib aḫmäd atïm имя мое Адиб Ахмед (Юг B_{469}); ratnapušpi atlïγ täŋri täŋrisi burχan atïn bu nom ẹrdïninij atï birlä ataju ïnča tep tezün пусть он, произнося имя будды – бога богов – по имени Ratnapuṣpa вместе с названием этой драгоценней сутры скажет следующее (*Uig* I 30_{1}); ata ornï atï oγulqa qalïr имя и положение отца достаются сыну (*QBN* 2012); 2. титул, звание: ẹki oγlïma jabγu šad at bertim двум моим сыновьям я дал титулы ябгу и шад (МЧ$_{19}$); atïm qul tapuγčï kör ornum qapuγ звание мое – раб-прислужник, место мое – у двери (*QBN* 53_{7}); 3. название, наименование: anuŋ bašï soγuqdïn ap aq turur anuŋ üčün anuŋ atï muz daγ turur от холода вершина ее (*горы*) вся белая, поэтому ее название – Ледяная гора (ЛОК 26_{7}); ädibniŋ jeri atï jügnäk ẹrür название родины (*букв*. местности) адиба – Югнак (Юг B493); kitab atï название книги (*QBK* 812).

AT II лошадь, конь: antaγ uluγ ölüg barγu tüšti kim jüklämäkkä kelṭürmäkkä at qaγaṭïr ud azlïq boldï столь много досталось добычи, что погрузить ее и привезти не хватало (*букв*. стало недостаточно) лошадей, мулов и быков (ЛОК 313); ẹr at ẹδärlädi мужчина оседлал коня (МК I 300); atïnda qodï tüšti он сошел с коня (*QBK* 19016).

AT- 1. бросать: kẹlsä kiši atma aŋar örtär külä если придет кто-либо, не бросай ему в лицо горячую золу (МК I 129); özüŋ otqa atma bu dünja ičün не бросайся в огонь ради благ мира (*QBK* 5814); ičindä tatïγ bolmasa ol qaγun anï tastïn atγu bolur если дыня невкусна внутри, то ее следует выбросить наружу (*QBK* 3049); 2. стрелять: ja qurup oq alïp adnaγu isig özintä adïrtim ... ẹrsär если я ... натянув лук,выстрелил и лишил кого-то (*букв*. другого) жизни (*Uig* II 8749); oqlarnï kökkäčä aṭuŋ пускайте стрелы до самого неба (ЛОК 394); 3. перен. отгонять, прогонять: mendä bulnur sẹvinč otï qaδγu atar есть у меня лекарство счастья, оно прогонит печаль (МК III 374).

AŠ[*ир.* āš?] I еда, пища: billglig kišilär bïsïγ jer ašïγ знающие (~ мудрые) люди едят пищу сваренной [до готовности] (QBN36$_1$); tonum qoj jüŋi tap jegüm arpa aš одежда моя – овечья шерсть, [этого] довольно, [а] пища моя – из ячменя (QBN343$_5$); süčig jaγlïγ aš сладкая, жирная пища(Suv591$_{22}$); 2. пир, угощенье; званый обед: qalï ašqa bęglär oqïsa sęni если беки позовут тебя на пир(QBK143$_{10}$); olardïn birisi küdänkä aš ol / ja sünnät asï ja toγursa oγul один из них (пиров) свадебный / или же по случаю обрезания или рождения сына (QBK27112,13).

AČ I голодный: kęčä jaṭza taŋta jana ač turur вечером ляжет [сытым], а утром снова голодный встает (*QBH* 103$_5$); jalïŋ boγazï todmaz ač jorïjur обнаженное (?) горло его не насыщается, [он] живет голодным (*TT* V115); atan jüki aš bolsa ačqa az körnür будет еды [на целый]вьюк верблюда – голодному [и это] покажется мало (МК 50$_{13}$).

AZ I 1. мало, немного: az udïŋ поспите немного (КР 55$_5$); biliglig eδi az biligsiz üküš [людей] сведущих (~ мудрых) очень мало, невежественных много(QBN 27$_4$); üküš az tep ajmas päzirlär deŋiz море не скажет, много или мало, [все] примет (Юг *B*$_{62}$); 2, редко: jaγïlïγ kišilär qutulmaqï az спасение (~ избавление) людей, имеющих врагов, [случается] редко(QBN 306$_{14}$); 3. немногочисленный; небольшой: az bodunïγ üküš qïltïm немногочисленный народ я сделал многочисленным (КТm$_{10}$); uluγ irkin azqïja ęrin tęzip bardï великий Иркин бежал с немногими только мужами (КТ$_{34}$); bu ïrq basïnta az ęmgäki bar в начале этого предзнаменования есть небольшая неприятность(ThSII$_{89}$); 4. в знач. сущ. малое количество: azïγ üküš qïltï малое он сделал большим (КТ$_{16}$); uquš azïn azlanma asγï üküš не гнушайся [и] малым количеством разума, [так как] пользы от него много(QBN34$_6$).

ЛИТЕРАТУРА

1. A.L.E. – Atlas Linguarum Europae sous la rédaction de A. Weijnen, rédacteur-en-chef Mario Alinei, Manuel Alvar, R.I.Avanesov et al. Première questionnaire (onomasiologie, vocabulaire fondamental). Secrétariat de la Rédaction de l'A.L.E., Nimègue 1973

2. Апресян Ю.Д. Экспериментальное исследование семантики русского глагола. М.: Наука, 1967.

3. Апресян Ю.Д. Лексическая семантика. М., 1968

4. Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л. и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003—2005. М.:Индрик, 2005, 193–214

5. Апресян Ю.Д., Дяченко П.В., Лазурский А.В., Цинман Л.Л. О компьютерном учебнике лексики русского языка // Русский язык в научном освещении. № 2 (14). 2007. С. 48–112.

6. Арутюнова Н.Д. К проблеме функциональных типов лексического значения // Аспекты семантических исследований. М.: Наука, 1980 . – С. 156–249. Позднее в сокращённом виде вошла в книгу: Арутюнова Н. Д. Язык и мир человека. М.: Языки русской культуры, 1999, с. 000–000?

7. БХРС – Большой хакасско-русский словарь / Под ред. О.В.Субраковой. Новосибирск, «Наука», 2006.

8. Виноградов В.В. Основные типы лексических значений слова // Вопросы языкознания», 1953, № 5.- Переизд.: Виноградов В.В. Избранные труды. Лексикология и лексикография. – М., 1977. – С. 162–189.

9. Грамматика хакасского языка / Под ред. Н.А.Баскакова. М., 1975.

10. Жолковский А.К., Леонтьева Н.Н., Мартемьянов Ю.С. О принципиальном использовании смысла при машинном переводе // Машинный перевод. М.: ИТМ и ВТ АН СССР, 1961. С. 17-46.

11. Кретов А.А. Анализ семантических помет в НКРЯ // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009, 240—257.

12. Кустова Г.И., Толдова С.Ю. НКРЯ: семантические фильтры для разрешения многозначности глаголов // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009, 258—276.

13. НКРЯ, http://www.ruscorpora.ru/corpora-sem.html

14. Розенцвейг В.Ю. Лексика имущественных отношений // Машинный перевод и прикладная лингвистика. Вып. 8. М., 1964.

15. СИГТЯ, Лексика – Сравнительно-историческая грамматика тюркских языков: Т. 4: Лексика. – М., 1997, 2001.

16. ЭСТЯ 2003 – Этимологический словарь тюркских языков: Общетюркские и межтюркские основы на буквы «Л», «М», «Н», «П», «С» / Авт. сл. статей Л.С. Левитская, Г.Ф. Благова, А.В. Дыбо, Д.М. Насилов, Е.А. Поцелуевский. М., 2003

17. Erdal 1991 – *Erdal M.* Old Turkic Word Formation. I-II. Wiesbaden, 1991.

18. Fillmore Ch. Frame semantics // Linguistic in the Morning Calm. Seoul – Hanshin 1982.

19. http://framenet.icsi.berkeley.edu/ framenet19.

MORPHOLOGICAL ANNOTATION SYSTEM FOR THE NATIONAL CORPUS OF THE CHUVASH LANGUAGE[1]

Pavel Zheltov

Chuvash State University, Cheboksary, Russia (428015, Cheboksary, Russia, Moskovskiy prospect 15), e-mail: chnk@mail.ru

The paper sets a task of development of morphological annotation system for the National Corpus of the Chuvash language. At present time the Turkic corpus linguistics is at early stage of working out. For the Chuvash language currently does not exist any special linguistic corpus.

Development of automatic natural language processing, the creation of national corpus of languages sets the task of developing specific standards for data representation. As a part of implementation of the system of morphological annotation for a corpus of the Chuvash language, the author considers linguistic traditions and norms reflected in the classic works on the morphology of Chuvash.

Создание национальных лингвистических корпусов, представляющих собой базу данных текстов на каком-либо языке, в которой каждое слово имеет морфологическую аннотацию с присущими ему морфологическими характеристиками, требует разработки системы соответствующих морфологических разметок.

При этом существующих сведений о морфологии и грамматике того или иного языка, изложенных в общепринятых грамматических справочниках, не всегда достаточно, т.к. при компьютерной обработке текстов, требующей полной формализации знаний о языке и его грамматике, часто обнаруживаются скрытые нюансы, не отраженные в классических лингвистических исследованиях.

Это в особенности касается тюркских языков, по которым до сих пор нет структурных грамматик, в которых эти языки исследовались бы с позиций их внутренних закономерностей, на основании которых и вырабатывались бы грамматические правила группировки и классификации тех или иных элементов.

Так, например, до сих пор нет ответа на вопрос: сколько падежей имеется в каждом из тюркских языков, равно как и не выработан критерий определения является или нет тот или иной элемент падежным аффиксом или служебным словом.

[1] Публикация подготовлена в рамках поддержанного РГНФ научного проекта № 15-04-00532.

Более того, имеются разные подходы к орфографии сложных слов относительно слитного/раздельного их написания.

Также следует отметить отсутствие единообразия при обозначении одних и тех же форм и категорий.

При этом задача осложняется еще и тем, что требуется придерживаться принятых для того или иного языка норм и правил, так как самоуправство в данной области может привести к тому, что обозначения, не согласованные с традиционными, будут непонятны языковедам и пользователям вообще, и таким национальным корпусом перестанут пользоваться.

В настоящее время тюркологами предпринимается попытка унификации аннотирования корпусов для тюркских языков в рамках научно-практического семинара «Унификация систем грамматической разметки в корпусах тюркских языков» (семинар UniTurk) и создания единого стандарта представления лингвистической информации.

Несомненно, что разработка подобного стандарта позволила бы организовать существующие и создающиеся корпуса тюркских языков в единое информационное пространство. Однако, изменение традиционных обозначений на новые требует переиздания грамматик и учебных пособий, переподготовку учителей и преподавателей национальных языков, на что требуется определенное время. Требуется также утверждение нового стандарта национальными комиссиями или иными законодательными органами, что часто может затягиваться из-за имеющихся разногласий внутри них, которые нередко основаны на личных амбициях и противоборстве лингвистических школ.

Поэтому в рамках создания системы морфологических разметок для национального корпуса чувашского языка была принята стратегия, основанная на установившихся к данному моменту традициях и нормах обозначений, общепринятых большинством чувашских грамматистов и языковедов и отраженная в классических трудах по чувашской морфологии.

Морфологическая разметка для национального корпуса чувашского языка имеет следующий вид:

N (noun) – существительное: ача «ребенок», Турă «Бог».

V (verb) – глагол (основа): тасат – «чистить», Çывар – «Спать».

Num (numeral) – числительное: пиллĕк «пять», çĕршер «по сто»;

Adv (adverb) – наречие: хăвăрт «быстро», малашне «впредь»;

Adj (adjective) – прилагательное: таса «чистый», лайăх «хороший»;

Pron (pronoun) – местоимение: кам «кто», никам «никто»;

Conj (conjuction) – союз: та, те «и», анчах «но», «однако»;

Part (practicle) – частица: ан, мар «не»;

Post (postposition) – послелог: витĕр «сквозь», пек «как»;

Onom (onomatopoetic) – подражательное слово: йăлтар-ялтар «подражание сверканию», сиянию, йăн-ян, «подражание звонку».

Имена существительные

2.1. Категория числа

Sing (singular) – единственное число: кĕнеке «книга», урапа «телега», «колесо»;

Pl (plural) – множественное число: кĕнекесем «книги», урапасем «телеги», «колесо»;

2.2. Категория падежа

Nom (nominative) – основной падеж: кĕнеке «книга»;

Gen (genitive) – притяжательный падеж: кĕнекен «книги»;

Dat-acc (dative) – дательно-винительный: кĕнеке «книге», «книгу»;

Loc (locative) – местный падеж: кĕнекере «в книге»;

Abl (ablative) – исходный падеж: кĕнекерен «из книги»;

Depr (deprivative) – лишительный падеж: кĕнекесĕр «без книги»;

Com (communative) – совместный падеж: кĕнекене, кĕнекепеле, кĕнекепелен «с книгой»;

Des (desiderative) – причинно-целевой падеж: кĕнекешĕн «ради книги», «из-за книги».

ЛИТЕРАТУРА

1. Желтов П.В. Лингвистические процессоры, формальные модели и методы: Теория и практика. Чебоксары: Изд-во Чуваш. ун-та, 2006.

2. Желтов П.В. Сопоставительно-сравнительное исследование морфем чувашского языка с применением формальных методов: диссертация … кандидата филологических наук. Чебоксары, 2010. 194 с.

3. Унификация систем грамматической разметки в корпусах тюркских языков (семинар UniTurk). Казань, 2014.

MORPHOLOGICAL TAGGING OF CRIMEAN TATAR ELECTRONIC CORPUS

Lenara Kubedinova[1], Ayrat Gatiatullin[2]

[1]Crimean Federal University, Simferopol, Crimea, Russia
[2]Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan Federal University

This article is devoted to the one of the urgent tasks for the Linguistic Corpus of the Crimean Tatar language – developing the system of morphological tags. In the context of the unification process of morphological categories of Turkic electronic corpora, that was initiated on the TEL–2014 conference, the comparison of the table of Tatar grammatical categories and the table of the Crimean Tatar affixes was conducted. Detailed analyses shows us a number of Crimean Tatar morphemes which are absent in the table for the Tatar language. The next step includes the comparison of these categories of the Crimean Tatar language with corresponding ones of the Turkish language.

Исследования в области корпусной лингвистики привлекают все больше и больше языков мира. Неоспоримым фактом является то, что объединение усилий ученых в исследованиях родственных языков, увеличивает скорость прогресса в познании возможностей языка и создании современных технологий для изучения и практического применения в прикладной лингвистике. Например, известны группы ученых занимающихся славянскими языками, германскими языками. В последние годы появляется много коллективов, создающих разработки и для тюркских языков. Такие коллективы есть в Турции (Стамбул), Азербайджане (Баку), Казахстане (Астана, Алмата), Китае (Урумчи), Башкортостане (Уфа), Чувашии (Чебоксары), Якутии (Якутск). В Татарстане разработками электронного корпуса татарского языка занимаются в научно-исследовательском институте «Прикладная семиотика» Академии наук Республики Татарстан.

В программе фундаментальных исследований президиума РАН отмечается, что «создание, развитие и использование электронных корпусов – это одно из наиболее передовых направлений современной лингвистики; именно в рамках этих направлений наиболее вероятны инновационные результаты как в области теоретической лингвистики (получение новых знаний об устройстве языка), так

и в области прикладной лингвистики (получение технологий нового поколения для автоматической обработки текстов и ускоренная модернизация методов лингвистических исследований)» [1]. Стремление взаимодействия с исследователями и разработчиками других тюркских корпусов реализовалось в проведении специального семинара Uniturk в рамках конференции по компьютерной и когнитивной лингвистике TEL–2014, на котором был начат процесс унификации морфологических категорий, которые используются при морфологической разметке тюркских электронных корпусов [2]. Данный семинар собрал ведущих специалистов, занимающихся разработкой и усовершенствованием электронных корпусов татарского, казахского, киргизского, башкирского, якутского, чувашского, хакасского, крымскотатарского и других тюркских языков.

В 2006 году был создан первый корпус электронных текстов крымскотатарского языка, авторами которого являются доцент кафедры английской филологии Таврической академии Крымского федерального университета Ленара Кубединова и научный сотрудник Института языкознания им. Л. Штура (Братислава, Словакия) Радован Гарабик. Лингвистический корпус крымскотатарского языка направлен на создание базы данных современного письменного языка.

Одной из актуальных задач для корпуса стала разработка программы морфологической разметки, для осуществления которой необходимо разработать систему морфологических тегов.

Нами было проведено сравнение таблицы грамматических категорий, представленных на конференции TEL – 2014, группой казанских разработчиков корпуса [2] и таблицы крымскотатарских аффиксов.

Анализ показал, что в таблице крымскотатарских морфем есть целый ряд морфем, которые отсутствуют в таблице для татарского языка.

Список этих морфем приведен в таблице 1.

Таблица 1

	Морфема	Алломорфы
1	-нен	
2	-джА	-джа, -дже, -ча, -че

3	-мАктА	-макъта/ -мекте
4	-Г[ъ]АйдЫ	-гъайды, -гейди, -къайды, -кейди
5	-мА+Й+ЫП	-майып, -мейип
6	-мАлЫ	-малы, -мели
7	-АрАк[ъ]	-аракъ, -ерек, -яракъ, -йерек
8	-мАдАн	-мадан, -меден
9	-ГЪАн+джА[къ]	-гъанджа(къ), -генджe(к), -къанджа(къ), -кендже(к)

Морфологические категории имени существительного в татарском и крымскотатарском языках в основном сопадают. Разница в именных категориях заключается в наличии в крымскотатарском языке инструментального падежа, который обозначается морфемой **-нен**. Исследователи крымскотатарского языка расходятся во мнениях по поводу инструментального падежа, одни ученые выделяют его, а другие считают его аффиксным послелогом. Кроме крымскотатарского инструментальный падеж выделяется еще в целом ряде тюркских языков, к числу которых относятся турецкий, казахский и чувашский языки.

Персональность (для личных форм глаголов и категории сказуемости существительных) в третьем лице единственном числе в крымскотатарском языке представлена аффиксами **-дыр/-дир**, **-тыр/-тир**, тогда как в татарском языке в данной категории аффиксы отсутствуют.

Этот аффикс также несет в себе и другие смысловые оттенки. Он, как и в татарском языке, относится к модальным аффиксам, выражающим сомнение (по грамматике А. Меметова). Например: *барадыр* «возможно ходит», *баргъандыр* «возможно сходил», *бараджакътыр* «возможно пойдет».

Что касается атрибутивных форм, производных от существительных, то четыре формы, представленные в татарском языке, имеют соответствия в крымскотатарском: атрибутив на **-лы** (мунитатив), атрибутив на **-сыз** (абессив), локативный атрибутив, генитивный атрибутив. Но, в крымскотатарском языке также существуют:

- суффиксы – **лыкъ, -лик, -лукъ, -люк** образуют имена прилагательные с основным значением предназначения: *къыш – къышлыкъ* (предназначенный на зиму (одежда, пища, дом)); *эки саатлыкъ иш = ике сәгатьлек эш* = работа, рассчитанная на два часа. В татарском языке этот аффикс также присутствует. Также возможны: *барганлык*, *барырлык* ‘способный пойти’, *барырлык җир* ‘место куда можно пойти’.

В крымскотатарском языке также существуют суффиксы прилагательных со значением недостаточности признака, неполного качества: **-тим, -(л)тым, -(юль)тим, -шын, -чыкъ, -чик**: *ешильтим боя* «зеленоватая краска», *сарылтым япракъ* «желтоватый листок», *къыскъачыкъ* «коротенький», *кучючик* «малюсенький». Но, данные аффиксы являются нерегулярными, так как присоединяются только к определенным словам, вследствие чего их лучше рассматривать в словаре.

Сравнительная категория прилагательного в крымскотатарском языке также выражается аффиксальным способом. Однако, для выражения этой категории в крымскотатарском языке используется морфема **–джА**, тогда как в татарском языке – морфема **–рАк**.

Наибольшие отличия между двумя исследуемыми языками мы наблюдаем в различных категориях глагола. Если мы говорим о временах глаголов, то в крымскотатарском языке существует длительное настоящее время, которое отсутствует в татарском языке, выражаемое суффиксами –**макъта/ -мекте**. Эта модель выражает действие, которое совершается продолжительно в настоящее время, оно началось до момента речи и продолжается поныне: *Динъленип бакътым, шытырдагъан шей чакъыл дегиль эди. Сес каналнынъ астында чыкъмакъта (Шамиль Алядин)* “Прислушался, то, что шуршало, не было галькой. Голос доносится со дна канала”.

Спряжение глаголов желательного наклонения в татарском языке совпадает со спряжением глаголов в императиве, тогда как в крымскотатарском языке глаголы желательного наклонения в прошедшем времени имеют свои аффиксы: **-гъайды, -гейди, -къайды, -кейди**. Эта форма выражает желание совершения или несовершения действия в широком плане: желать что-то делать; чтобы что-то свершилось; действие, которое свершилось бы при определенном условии: *Меним не къабаатым бар, тиймегейди, тиймез*

эдим (Юсуф Болат). “В чем моя вина, если бы он не тронул меня, то бы и я его не тронул”.

Деепричастия в крымскотатарском языке также имеют отличные от татарского языка морфемы. Если положительная форма деепричастия сопутствующего действия в крымскотатарском и татарском языках совпадают, то отрицательная форма того же деепричастия представлена разными морфемами: **-мА+Й+Ып** и **-мА+Й+чА** соответственно: *алмайып* «не беря», *кельмейип* «не приходя», *окъумайып* «не читая», *ишлемейип* «не работая».

Деепричастие с предшествующим значением, т.е. выражающее действие, после которого моментально совершается другое действие, имеет морфему **-АрАкъ**: *аларакъ* «как только он взял», *кетерек* «как только он ушел», *башлаякъ* «как только он начал».

Крымскотатарское деепричастие с противоположным действием, т.е. обозначающее предел главного действия во времени или указывающее на момент совершения главного действия, имеет морфему –ГЪАнджА[къ]: *алгъанджакъ* «когда он брал», *бергенджек* «когда он давал», *айткъанджакъ* «когда он сказал», *кеткенджек* «когда он ушел».

Если обратиться к модальным формам глагола, то в татарском языке наблюдается большее разнообразие форм, чем в крымскотатарском языке. Морфемы условной модальности (conditional) в обоих языках совпадают: **-сА**.

Категорию должествования в крымскотатарском языке можно выразить двумя способами **-магъа керек**, и **-малы**, аналогом которым в татарском является **-ырга кирәк**. Что касается других категорий (форма, выражающая значение возможности; форма, выражающая значение намерения (дезидератив); форма, выражающая значение предостережения), то они не представлены в крымскотатарском языке.

Следует отметить категории татарского языка, которые отсутствуют в крымскотатарском языке:

1) форма причастия, выражающая значение регулярно совершаемого действия, которое характеризует субъект;

2) инфинитив на **–ырга** (в крымскотатарском языке только одна инфинитивная форма);

3) способы глагольного действия – раритивы **–ГалА и -ЫштЫр**;

4) модальные формы глагола – форма, выражающая значение возможности; форма, выражающая значение намерения (дезидератив); форма, выражающая значение предостережения;

5) собирательное числительное;

6) модальные средства выражения вопросительности на **-мыни**, неопределенности, предостережения – три формы, выражающие значение уподобления.

7) следует отметить, что форма (на **-дай)**, выражающая значение уподобления, хотя не входит в рамки литературного крымскотатарского языка, но активно используется и в современной поэзии.

По результатам сравнительного анализа была построена таблица, в которой приводятся татарские аналоги для аффиксов крымскотатарского языка.

Таблица 2

	Крымскотатарская морфема	Аналог в татарском языке	Пример
1.	-нен		
2.	-джА	-рАк	узунджа ‘длиннее’
3.	-мАктА		бармакта ‘сейчас идет’
4.	-Г[ъ]АйдЫ	-сА иде	баргъайдым ‘пойти бы мне тогда’
5.	-мА+Й+Ып	-мА+Й+чА	бармайып ‘не сходив’
6.	-мАлЫ	-ЫргА тиеш	бармалы ‘должен пойти’
7.	-АрАк[ъ]	-ГАн-ДА	бараракъ ‘как только он пошел’
8.	-ГЪАн+джАкъ		баргъанджакъ ‘когда он пошел’
9	-мАдАн	-мЫйчА	бармадан ‘еще не сходив’

Заключение

Проведенный сравнительный анализ показал наличие в крымскотатарском языке целого ряда аффиксов, отсутствующих в татарском языке. Это показывает необходимость введения системы обозначений для таких категорий. С целью введения обозначений этих категорий ведется изучение разработок для турецкого языка,

поскольку велика вероятность того, что для многих из этих категорий специалистами по огузской подгруппе тюркских языков уже используется система обозначений морфологических категорий.

ЛИТЕРАТУРА

1. Труды казанской школы по компьютерной и когнитивной лингвистике TEL-2012. – Казань: Изд-во «Фэн» Академии наук РТ, 2012. – 176 с.

2. Труды казанской школы по компьютерной и когнитивной лингвистике TEL-2014. – Казань: Изд-во «Фэн» Академии наук РТ, 2014. – 298 с.

3. Меметов А.М. Земаневий къырымтатар тили. – Симферополь: Къырым девлет окъув педагогика нешрияты, 2006. – 320 с. – Къырымтатар тилинде.

4. Меметов А., Мусаев К. Крымтатарский язык. Ч. I. Общие сведения о языке. Ч. II. Морфология. Учебное пособие. – Симферополь: Крымское учебно-педагогическое государственное издательство, 2003. – 288 с. – На русском языке.

SYNTACTIC ANNOTATION OF KAZAKH: FOLLOWING THE UNIVERSAL DEPENDENCIES GUIDELINES. A REPORT

Aibek Makazhanov[1] , Aitolkyn Sultangazina, Olzhas Makhambetov and Zhandos Yessenbayev

National Laboratory Astana, 53 Kabanbay Batyr ave.,
Astana, 010000, Kazakhstan
[1]aibek.makazhanov@nu.edu.kz

The present work is a report on the authors' first attempt to use the universal dependencies (UD) (de Marneffe et al., 2014) standard for syntactic annotation of Kazakh. The report is a result of a manual annotation of 300 sentences randomly chosen from the Kazakh Language Corpus (Makhambetov et al., 2013). We focus primarily on providing an extensive list of annotation examples that covers syntactic relations currently used in the framework of the UD project (ud Documentation, 2015). We also report on certain language-specific issues that may require special treatment.

1. Introduction

A computational processing of Kazakh, a Turkic language of the Qypchaq group, has received a fair amount of attention in the recent years. Most notable research efforts were concentrated on building language resources (Makhambetov et al., 2013; Altenbek and Xiao-long, 2010) and morphological processing (Kessikbayeva and Cicekli, 2014; Washington et al., 2014; Makhambetov et al., 2015) with related applications in POS tagging (Makazhanov et al., 2014a) and spelling correction (Makazhanov et al., 2014b). Sadly, existing works on language resources for Kazakh do not provide a detailed coverage of the process of syntactic annotation. In this paper we try to fill the gap starting with the very basics. Namely, we provide an extensive list of examples of annotating various syntactic structures following the UD standard. The standard dictates rules, which are generic enough to be applied to (ideally) any language and, at the same time, flexible enough to account for most of language specific exceptions. While the standard provides detailed guidelines to all major aspects of annotation, i.e. tokenization, POS, morphology and syntax, in the present work we focus only on the latter.

2. Syntactic relations

As its name suggest, the universal dependencies standard represents syntactic structures in a form of typed dependency relations between words. Each word is either the dependent of one other word (its governor) in a sentence or of a notional ROOT of a sentence (ud Documentation, 2015).

This section conveys the essence of the present work. In the following nine subsections we provide annotation examples for 40 relations 38 of which are present in the universal dependencies documentation (UDD) for English. Each subsection corresponds to a category of relations per UDD classification. Definitions of most of the relations are given as they appear in UDD with minor editions. Throughout the paper we refer to specific relations using `R(d,g)` notation, where `R` is the relation name, and `d` and `g` correspond to the dependent and the governor respectively.

2.1. Core dependents of clausal predicates

Nominal subject (nsubj)

A nominal subject is a noun phrase (or other nominal) which is the syntactic subject of a clause, e.g. nsubj (3,4):

```
мұнда(1) үлкен(2)  шара(3)     өтеді(4)
 here      big     event   will take place
```

The governor of a subject might not always be a verb: non-verbal predicates are possible in cases of a copular predication. If copula is overt, the root of the clause is the complement of the copular verb. The two examples below illustrate cases with covert and overt copulas respectively:

```
      сары (1) жылы(2) түс(3)
      yellow   warm   color
жеті (1) жұп(2) па(3) тақ(4) па(5) еді(6)
 seven    even   .     odd    .     was
```

In the first example we have a noun predicate `түс`: nsubj(1,3). In the second example we have two adjectival predicates `жұп` and `тақ`: nsubj(1,2) and nsubj(1,4). Also, in the first example the subject is an adjective and in the second – numeral(s).

Passive nominal subject (nsubjpass)

A passive nominal subject is a noun phrase (NP) which is the syntactic subject of a passive clause, e.g. nsubjpass(3,4):

```
гүр(1) еткен(2) дауыс(3) естілді(4)
 roar   making   voice   was heard
```

Direct object (dobj)

This relation holds between a verbal phrase (VP) and its accusative object, with a VP being the governor, e.g. dobj(4,5):

```
бұл(1) көрсеткіш(2) 75(3) пайызды(4)  құрайды(5)
 this  indicator     75   percent   makes up for
```

Indirect object (iobj)

Same as the direct object relation, but with a non-accusative object, e.g. iobj(2,4):

```
төрт(1) екіге(2)    қалдықсыз(3)     бөлінеді(4)
 four   to two  without a remainder is divisible
```

Clausal subject (csubj)

A clausal subject is a clausal syntactic subject of a clause, i.e., the subject is itself a clause. As in the case with a nominal subject, the governor of this relation might not always be a verb, e.g. csubj(3,2):

```
сен(1) ол(2) өтірікші(3)
 you    it      liar
```

In the example above (1)-(2) makes up a clause (literally: you are he/she/it) which depends on (3). The whole sentence translates as "*you are the one who is a liar*".

Clausal passive subject (csubjpass)

A clausal passive subject is a clausal syntactic subject of a passive clause, e.g. csubjpass(2,4):

```
қисап(1)  құратындар(2)  да(3) табылатын(4)
 count   those who make   too   were found
```

Clausal complement (ccomp)

A clausal complement of a verb or adjective is a dependent clause with an internal subject which functions like an object of the verb or adjective, e.g. ccomp(2,4):

```
сайттың(1) маңызды(2) екенін(3) түсінбей(4) жатыр(5)
 web site  important   being      don't understand
```

Open clausal complement (xcomp)

An open clausal complement (xcomp) of a verb or an adjective is a predicative or clausal complement without its own subject. The reference of the subject is necessarily determined by an argument external to the xcomp (normally by the object of the next higher clause, if there is one, or else by the subject of the next higher clause, e.g. xcomp(4,5):

```
тұлға(1) кедендік(2) рәсімді(3) өзгертуге(4)құқылы(5)
person   customs    procedure    change     eligible
```

2.2. Non-core dependents of clausal predicates

Noun phrase as adverbial modifier (nmod:npmod)

This is a subtype of a nominal modifier (see nmod) relation, which is used when something syntactically NP is used as an adverbial modifier, e.g. nmod:npmod(2,3):

```
Асылдың(1)  қасында(2)  болғаныңыз(3) дұрыс(4)
 Asyl.NNP  by her side   your being    right
```

Temporal modifier (nmod:tmod)

This relation marks cases when something syntactically NP is used as a temporal modifier, e.g. nmod:tmod(3,5):

```
қызметін(1) 1933(2) жылы(3) ғана(4) бастаған(5)
his service  1933    year    only     began
```

Adverbial modifier (advmod)

An adverbial modifier of a word is a (non-clausal) adverb or adverbial phrase that serves to modify the meaning of the word, e.g. advmod(2,3):

```
Бақытқұл(1)   мүлдем(2)  өзгерді(3)
Baqytkul.NNP  completely  changed
```

Adverbial clause modifier (advcl)

An adverbial clause modifier is a clause which modifies a predicate, as a modifier not as a core complement. This includes things such as a temporal clause, consequence, conditional clause, purpose clause, etc.

In contrast to the advmod relation the modifier must be clausal, e.g. advcl(3,5):

```
кеуденде(1) намыс(2)  болса(3)  , (4) қимылда(5)
   in your chest  pride   if exists  ,       act
```

2.3. Special clausal dependents

Vocative (vocative)

This relation is used to mark addresses in direct speech. It links the addressee's name to the predicate of its host sentence, e.g. vocative(2,5):

```
хан(1)  ием(2) ,(3) Жүзіқараны(4) алып(5) келдік(6)
 khan  my master ,   Zhuziqara.NNP  we have brought
```

Discourse element (discourse)

This relation is used for interjections and other discourse particles and elements. As in the vocative relation the head of a discourse element is attached to the main predicate of a corresponding sentence, e.g. discourse(1,4):

```
жоқ(1) ,(2)  хат(3)   жазып(4) берейін(5)
  no   ,     letter      let me write
```

Auxiliary (aux)

An auxiliary of a clause is a non-main verb of the clause. In the following example, an auxiliary (4) attaches to the main verb (3) to make a present continuous tense verb, e.g. aux(4,3):

```
жағдайдың(1)  барлығын(2) жасап(3) отыр(4)
of conditions     all         is making
```

Copula (cop)

A copula is a relation between a copular verb and its complement. We consider the verbs "*е*" and "*бол*" (correspond to *be*) and possessive "*бар*" and "*жоқ*" (correspond to *[not] have*) as copular verbs. Per UDD copular words are attached to their complements, e.g. cop(3,2) and cop(4,3):

```
        менде(1) су(2) бар(3)
          on me   water  have
Өтемістен(1) туған(2) он(3)  едік(4)
from Otemis   born     ten   we were
```

Marker (mark)

A marker is the word introducing a clause subordinate to another clause. The mark is a dependent of the subordinate clause head, e.g. mark(3,1) and mark(2,1):

```
егер(1) сүт(2)  болмаса(3) ,(4) ештеңе(5)   алма(6)
 if     milk  there is no ,    anything  do not buy
келгеннен(1) кейін(2) ,(3) бірден(4)     жатты(5)
from coming   after   ,     at once    went to bed
```

Punctuation (punct)

Punctuation is attached to the head of corresponding clause, e.g. punct(3,1) and punct(6,5):

```
келгеннен(1) кейін(2) ,(3) бірден(4) жатты(5)   .(6)
from coming   after   ,    at once went to bed .
```

2.4. Noun dependents

Numeric modifier (nummod)

A numeric modifier of a noun is any number phrase that serves to modify the meaning of the noun with a quantity, e.g. nummod(3,4):

```
Баба-атаға(1) да(2)  үш(3) күн(4)  түнеді(5)
To Baba-ata   also  three   day   slept-over
```

Note that ordinal numerals are treated as adjectival modifiers, as they express property rather than quantity.

Appositional modifier (appos)

An appositional modifier of an NP is an NP immediately to the right of the first NP that serves to define or modify that NP. Its head is governed by the head of the modified NP e.g. appos(5,2):

```
Жаннет(1)   Чочаева(2) , (3) байқауға(4)  қатысушы(5)
Zhannet.NNP Chochayeva.NNP , to contest  who attends
```

Nominal possessive modifier (nmod:poss)

This relation is used to mark constructions with definite or indefinite genitives that depend on possessive nominals, e.g. nmod:poss(1,2) (definite) and nmod:poss(2,3) (indefinite).

```
Асылдың(1)  қасында(2)  болғаныңыз(3) дұрыс(4)
 Asyl.GEN   near.P3SG    your being    right
```

```
тартылған(1) инвестиция(2)   көлемі(3)
 involved     investments   volume.P3SG
```

Relative clause modifier (acl:relcl)

This relation is used when a relative clause is modifying a noun. The head of the clause is governed by the noun it modifies, e.g. acl:relcl(2,3) and acl:relcl(4,5):

```
кейіннен(1)   алынған(2)    қаражат(3)
  lately    were received    funds
жас(1) ,(2) еті(3) тірі(4)  жігіт(5)
young  ,     meat   alive    guy
```

Determiner (det)

A relation that is used to mark determiner-NP attachments. We consider demonstrative pronouns and certain indefinite pronouns as determiners, e.g. det(1,2) and det(3,4):

```
барлық(1) шығыстарды(2) бұл(3) одақ(4) көтереді(5)
    all      expenses     this   union    lifts
```

Adjectival modifier (amod)

An adjectival modifier of an NP is any adjectival phrase that serves to modify the meaning of the NP, e.g. amod(1,2):

```
электронды(1) тасығышта(2) жасалған(3)  ақпарат(4)
 electric     on carrier     made      information
```

We treat ordinal numerals as amod as well, e.g. amod(2,3):

```
қызметін(1) 1933(2) жылы(3) ғана(4) бастаған(5)
his service  1933    year    only    he began
```

2.5. Compounding and unanalyzed

Compound (compound)

This relation is used to link compound nouns, numerals, and superlative forms of adjectives and adverbs. In all cases elements are connected in a flat head-last fashion, i.e. the rightmost elements governs all the rest (no chain structure), e.g. compound(1,2) and compound(1,3) [the first example below]; compound(1,2) [the second and third examples]:

```
            қозы(1)     жауырын(2)    жебе(3)
             lamb    shoulder blade    arrow
он(1)      екі(2) орындық(3)| ең(1)  жүйрік(2) ат(3)
 ten [twelve] two    chairs  | the most fast    horse
```

Light verb construction (compound:lvc)

This relation is used to accommodate so called light verb constructions, i.e. nominal-verb constructions where a verb loses its direct meaning and behaves like an "auxiliary" of a nominal to produce a new meaning. Such constructions are also annotated in a flat head-last fashion. In the following example "to pull cigarettes" compound:lvc(1,2), and "to pull guitar" compound:lvc(4,5) are literal translations of "smoke cigarettes" and "play guitar" respectively:

```
темекі(1)    тартқанша(2)     ,(3) гитара(4) тарт(5)
cigarettes instead of pulling ,  guitar   pull
```

Figure of speech (compound:fos)

This relation is used to link (adjacent) words, which as a whole comprise a figure of speech. The reason we treat such cases separately is that many of such constructions have a very common syntax. Compare: `шашы(hair.P3SG) ұзын(long) қыз(girl)` and `басы(head.3SG) бос(free) қыз(girl)`. While the first phrase is transparent (*a girl with long hair*), the second translates as *a girl who is not married* (literally: *a girl with a free head*). As in the case with all other compounds, this relation is annotated in a flat head-last structure, e.g. compound:fos(1,2):

```
      еті(1)           тірі(2)  жігіт(3)
       meat  [lively]  alive     guy
```

Name (name)

This relation is used for proper nouns constituted of multiple nominal elements. In general, names are annotated in a flat, head-last structure, in which all words in the name modify the first one using the name label, e.g. name(1,2) and name(1,3):

```
      Геннадий(1) Геннадиевич(2) Головкин(3)
       Gennady     Gennadievich   Golovkin
```

For organization names with clear syntactic modification structure, the dependencies should reflect the syntactic modification structure using regular syntactic relation, e.g acl(1,2) and nmod:poss(2,3):

```
Біріккен(1) Ұлттар(2)   Ұйымы(3)
  United     Nations  Organization
```

Multi-word expression (mwe)

This relation is used for certain fixed grammaticized expressions that behave like function words or short adverbials. As in the case with all other compounds, this relation is annotated in a flat head-last structure, e.g. mwe(1,2) [then+too=anyway] and mwe(5,6) [but+too=but/however]:

```
сонда(1) да(2) бар(3) ,(4) бірақ(5) та(6)  байқа(7)
 then    too   go    ,      but     too  be careful
```

Foreign words (foreign)

The relation is used to label sequences of foreign words. These are given a linear analysis: the head is the first token in the foreign phrase. The relation does not apply to loanwords or to foreign names. It applies to quoted foreign text incorporated in a sentence/discourse of the host language, e.g. foreign(3,2) and csubj(1,2):

```
  Ол(1)      само(2)       собой(3)
it[KK]   itself[RU]   by itself[RU]
```

Here the Russian phrase roughly corresponds to "*self-explanatory*", and the whole sentence translates as "*it is self-explanatory*".

Goes with (goeswith)

This relation links two parts of a word that are separated in text that is not well edited. The head is in some sense the "main" part, often the second part, e.g. goeswith(2,3):

```
отырыс(1) Қазақстан(2)  да(3)   өтуі(4)   тиіс(5)
 meeting   Kazakhstan +  in   take place  should
```

2.6. Coordination

Conjunct (conj)

A conjunct is the relation between two elements connected by a coordinating conjunction and/or punctuation. The head of the relation is

the last conjunct and all the other conjuncts depend on it via the conj relation, e.g. conj(1,5) and conj(3,5):

```
нан(1) ,(2) сүт(3) және(4) майды(5) сатып(6) ал(7)
bread  ,     milk   and    butter        buy
```

Coordinate clauses are treated the same way as coordination of other constituent types, e.g. conj(2,7) and conj(4,7):

```
басын(1) изеп(2) ,(3) жымиды(4) да(5) қолын(6) берді(7)
  head    nod    ,      smile    and    hand     gave
```

Coordinating conjunction (cc)

This is a relation between the main (the last in our case) conjunct and the coordinating conjunction delimiting another conjunct. In the previous example the relation holds between (5) and (7), i.e. cc(5,7).

2.7. Case-marking, prepositions, possessive

Case marking (case)

In Kazakh there is a group of postpositions that attach to nouns marked for a certain case (usually nominal and dative). We treat such postpositions as noun dependents, linking them in a similar manner, e.g. case(3,2):

```
инфекция(1) ауа(2) арқылы(3)   таралады(4)
 infection   air      by     is distributed
```

2.8. Loose joining relations

List (list)

The list relation is used for chains of comparable items. All items of the list shoud modify the first one, e.g. list(2,1) and list(4,1):

```
автор(1) тел.(2) : 000(3) эл.пошта(4) : bir@eu.kz(5)
 author   tel.   :  .       e-mail    :     .
```

Dislocated elements (dislocated)

The relation is used for fronted or postposed elements that grammatically do not fit in anywhere in a sentence. Dislocated elements are attached to the same governor as the dependent that they double for, e.g. dislocated(3,6) and dislocated(6,2):

```
 сен(1) және(2) мен(3) ,(4) офис(5) біздікі(6)
  you     and     I     ,     ofice    of us
```

```
мынадан(1) ұрттап(2) көр(3) :(4) тұзды(5) шай(6)
   this       take a sip    :     salty    tea
```

Parataxis (parataxis)

Parataxis (from Greek for "place side by side") occurs between the main verb of a clause and other sentential elements, such as a sentential parenthetical, a clause after a ":" or a ";", or two sentences placed side by side without any explicit coordination or subordination. Annotation is flat, e.g. parataxis(2,4) and parataxis(2,6):

```
ешқайда(1)  шықпайды(2),жабық(3)  жерде(4),арақ(5) ішпейді(6)
 nowhere     go out    , closed   in place ,  vodka  not drink
```

Remnant in ellipsis (remnant)

The relation is used to provide a satisfactory treatment of ellipsis (in the case of gapping and stripping, where a predicational or verbal head gets elided) without having to postulate empty nodes in the basic representation. Unlike for conj relation, remnant uses a chaining analysis where each subsequent remnant depends on the immediately preceding remnant/correlate, e.g. remnant(3,1)→remnant(5,3) and remnant(4,2)→remnant(6,4):

```
Мен(1) үй(2) ,ағам(3)  күн(4), апам(5) гүл(6) салды(7)
 I   house  ,my brother  sun,my sister flower  drew
```

Instances of stripping typically occur when there is only one argument in the second clause, but with an accompanying adverbial modifier such as *not* or *only*, e.g. remnant(4,1):

```
Ешкім(1) келген(2) жоқ(3) , мен(4) ғана(5)
 Nobody    came     not   ,   I     only
```

2.9. Others

Root (root)

The relation points to the root of the sentence. A fake node "ROOT" is used as the governor. The ROOT node is indexed with "0", since the indexation of real words in the sentence starts at 1, e.g. root(3,0)→csubj(2,3)→subj(1,2):

```
ROOT(0) сен(1) ол(2) өтірікші(3)
   .      you    it      liar
```

Unspecified dependency (dep)

A dependency is labeled as *dep* when a system is unable to determine a more precise dependency relation between two words. This may be because of a weird grammatical construction, a limitation in software, a parser error, or because of an unresolved long distance dependency.

3. Discussion

We did not specify the negation (*neg*) relation, because under the current scheme analytic negation markers, i.e. words *жоқ* and *емес*, are classified as copulas. Furthermore, both can be inflected (e.g. *қорқақ*[coward.nom] *емеспін*[not.A1sg]; *айтқан*[tell.PTCP] *жоқсын*[not.A1sg]), which is not a typical behavior of a negation particle. A compromise might be to introduce a sub-relation *cop:neg*.

We did not specify the clausal noun modifier relation (*acl*), because, as much as we tried, we could not find examples where such clauses were not relative (*acl:relcl*).

Due to morphological derivation, in some cases a dependent's attachment to its governor word makes sense only if done to a part of the word and not the whole. Consider: *үлкен*[big.ADJ] *үйдегілер*[house.ATTR.PL] (those in a big house). If we attach *үлкен* to *үйдегілер* directly, the adjective will be modifying *those in a house*, not the *house*, i.e. *those in a house are big*. Although the UD scheme allows defining multi-token words, such cases must be annotated consistently, and therefore tokenization convention must be agreed upon.

4. Conclusion

We have provided an extensive list of syntactic annotation examples for Kazakh. Annotation follows the universal dependency standard. The list covers a total of 40 dependencies. We have also discussed certain cases which we had difficulty to annotate using the existing UD relations. In the future we plan to accommodate those cases by defining sub-relations that should account for their specifics.

Acknowledgements

This work has been funded by the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan.

We also would like to thank Francis M. Tyers and Jonathan N. Washington for their kind advice.

REFERENCES

Gulila Altenbek and Wang Xiao-long. (2010). Kazakh segmentation system of inflectional affixes. In Joint Conference on Chinese Language Processing (pp 183–190). CIPS-SIGHAN.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. (2014). Universal Stanford dependencies: A cross-linguistic typology. In Nicoletta Calzolari et al. (Eds.), LREC (pp 4585–4592). ELRA.

Gulshat Kessikbayeva and Ilyas Cicekli. (2014). Rule based morphological analyzer of Kazakh language. In Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM (pp 46–54). Baltimore, Maryland: ACL.

Aibek Makazhanov, Olzhas Makhambetov, Islam Sabyrgaliyev, and Zhandos Yessenbayev. (2014a). Spelling correction for Kazakh. In Proceedings of the 2014 Computational Linguistics and Intelligent Text Processing (pp 533–541). Kathmandu, Nepal: Springer Berlin Heidelberg.

Aibek Makazhanov, Zhandos Yessenbayev, Islam Sabyrgaliyev, Anuar Sharafudinov, and Olzhas Makhambetov. (2014b). On certain aspects of Kazakh part-of-speech tagging. In Proceedings of the Application of Information and Communication Technologies (pp 1–4). IEEE.

Olzhas Makhambetov, Aibek Makazhanov, Zhandos Yessenbayev, Bakhyt Matkarimov, Islam Sabyrgaliyev, and Anuar Sharafudinov. (2013). Assembling the Kazakh language corpus. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp 1022–1031). Seattle, Washington: ACL.

Olzhas Makhambetov, Aibek Makazhanov, Islam Sabyrgaliyev, and Zhandos Yessenbayev. (2015). Data-driven morphological analysis and disambiguation for Kazakh. In Proceedings of the 2015 Computational Linguistics and Intelligent Text Processing (pp 151–163). Cairo, Egypt: Springer.

ud Documentation. (2015). http://universaldependencies.github.io/docs/. Accessed: 2015-06-19.

Jonathan Washington, Ilnar Salimzyanov, and Francis Tyers. 2014. Finite-state morphological transducers for three Kypchak languages. In Nicoletta Calzolari et al. (Eds.), LREC (pp 3378–3385), Reykjavik, Iceland: ELRA.

DEVELOPMENT OF SEMANTIC MARK-UP FOR THE CORPUS OF TUVAN LANGUAGE[1]

Baylak Oorzhak,
Arzhaana Khertek [2]

Tuvan State University
Kyzyl, Tuva, Russia

The article presents the development of a semantic annotation of texts for the corpus of Tuvan language, which would allow working with fragments of text required to be included in the educational-methodical complexes of the Tuva language for primary and General secondary education. The proposed semantic markup is developed based on semantic bits of words that are marked lexico-semantic labels for Tuvan and Russian languages. The tokens of the Tuvan language are divided into four basic semantic categories: Person, Animal, Subject, Natural objects and phenomena that permeate virtually the entire system vocabulary. Then they are broken down into smaller semantic subclasses.

The proposed semantic markup will be used later in the semantic processing of texts.

В настоящее время продолжается работа над разработкой Электронного корпуса текстов тувинского языка (ЭКТТЯ), начатого в 2011 г. при поддержке гранта Российского гуманитарного научного фонда. На сегодняшний день электронный корпус тувинского языка включает в себя тексты тувинской художественной литературы и фольклора разных жанров, словари, а также базы данных по именным и глагольным основам.

---

[1] «Работа ведется на средства гранта РГНФ № 15-04-12030 «Система автоматического морфологического и синтаксического анализа для корпусов миноритарных тюркских языков России»».

[2] oorzhak.baylak@mail.ru, khertek-ab@yandex.ru_

На основе «Морфемно-орфографического словаря тувинского языка» и баз данных по именным и глагольным основам была разработана морфологическая разметка корпуса, которая позволяет производить автоматический поиск морфем в заданном тексте и используется специалистами в исследованиях грамматических явлений тувинского языка. Нужно сказать, что существующая программа обеспечения методов поиска на данный момент находится на стадии совершенствования.

Наряду с этим возникла необходимость в разработке семантической разметки в связи с работой над созданием учебно-методических комплексов по тувинскому языку для начального и общего среднего образования. Морфологическая разметка позволяет работать над грамматическим материалом, что, безусловно, необходимо при создании школьных учебников. Однако, основное выдвигаемое требование ФГОС к УМК по языковым дисциплинам – формирование коммуникативной компетенции – невозможно без правильно подобранного лексического материала в учебнике.

Разрабатываемая семантическая разметка позволит оперировать с фрагментами текста, необходимого для включения в содержание учебника, на основе семантических разрядов слов и обслуживающих их лексико-семантических помет. Лексемы тувинского языка условно выделены в четыре базовые семантические категории: Человек, Животное, Предмет, Природные объекты и явления, которые пронизывают систему лексики. Далее они разбиваются на более дробные семантические подклассы.

Имя существительное

Лексико-семантические блоки:

- Предметные имена: человек
  животное
  предмет
  природные объекты и явления
- Непредметные имена: абстрактные понятия
- Имена собственные: человек, местность.

Лексико-семантические пометы:
Предметные имена

Кижи/Человек	***Төрел-аймак аттары***/ Имена родства	*авай* ‘мама’, *угбай* ‘сестра’, *даай* ‘дядя’	
	Профессия	*эмчи* ‘врач’, *башкы* ‘учитель’, *ыраажы* ‘певец’	
	Этноним	*кыдат* ‘китаец’, *бурят* ‘бурят’, *моол* ‘монгол’	
Дириг амытан/ Животное	***Дириг амытан***/ Животные	***черлик***/дикие	*адыг* ‘медведь’, *дииң* ‘белка’
		азырал/ домашние	*инек* ‘корова’, *ыт* ‘собака’
	Куштар/Птицы	***черлик***/дикие	*хартыга* ‘коршун’, ус-*кушкаш* ‘ремез’
		азырал/ домашние	*дагаа* ‘курица’, *кас* ‘гусь’
	Балыктар/Рыбы	*ак-балык* ‘елец’, *шортан* ‘щука’	
	Курт аймаа/ Насекомые	*шартылаа* ‘кузнечик’, *ары* ‘пчела’	
Чүүл/Предмет	***Бүдүмелдер***/ Вещества и материалы	*суг* ‘вода’, *чугай* ‘известь’, *торгу* ‘шелк’	
	Эт-херексел/Быт	*аптара* ‘сундук’, *хеп* ‘одежда’, *аяк-сава* ‘посуда’	
	Бажың-балгат, тудуг объектилери/ Здания и сооружения	*көвүрүг* ‘мост’, бажың ‘дом, здание’	
	Эдилел/инструменты	*балды* ‘топор’, *сыырткыыш* ‘удочка’	

Бойдус объектилери болгаш бойдустуң болуушкуннары/ Природные объекты и явления	***Үнүш***/растения	*оът-сиген* ‘трава’, *ыяш* ‘дерево’
	Агаар байдалы/ Погодные явления	*чаъс* ‘дождь’, *кызаңнаашкын* ‘гроза’, *челээш* ‘радуга’
	Дээр объектилери/ небесные тела	*хүн* ‘солнце’, *сылдыс* ‘звезда’
	Ландшафт	*даг* ‘гора’, *хем* ‘река’, *хову* ‘степь’

Непредметные имена

Туугай билиишкиннер/ Абстрактные понятия	***Сагыш-сеткил илерээшкини***/ Эмоции	*өөрүшкү* ‘радость’, *муңгарал* ‘горе’
	Миннишкин/ Чувственное восприятие	*дааш* ‘шум’, *амдан* ‘вкус’, *өң* ‘цвет’
	Ниити билиишкиннер/ Универсальные представления	*болуушкун* ‘событие’, *кылдыныг* ‘действие’, *байдал* ‘обстоятельство’, *үе* ‘время’

Имена собственные

Кижи/ Человек	***Ат***/Имя	*Чечек*, *Артыш*, *Менги, Кара-кыс*
	Адазының ады/ Отчество	*Дүрген-оолович*, *Бай-Караевна*
	Фамилиялар/ Фамилии	*Сарыг-оол, Сотпа*
	Аймак ады/ Названия родов	*Хертек*, *Ооржак*, *Тюлюш*
Дириг амытан/ Животное	***Аът***/Лошадь	*Калчан-Шилги, Сарала.*
	Инек/Корова	*Дөңгүр, Дагыр-Мыйыс*
	Ыт/Собака	*Ак-Төш, Көстүк, Калдарак.*
Черлер/ Местность	***Черлер аттары***/ Топонимы	*Кызыл*, *Чаа-Хөл, Кунгуртуг*

Имя прилагательное

Разряды:

- Качественные
- Относительные

Лексико-семантические пометы:
Качественные

Кижи/Человек	физические качества	*мөге* ‘сильный’, *семис* ‘толстый’, *бедик* ‘высокий’.
	умственные/ психические качества	*чазык* ‘приветливый’, *угаанныг* ‘умный’, *дидим* ‘смелый’.
Дириг амытан/ Животное	физические качества	*эъткир* ‘упитанный, мясной’, *сүткүр* ‘молочный’.
	масть	*хоор* ‘каурый’, *ала* ‘пегий’,
Бойдус объектилери болгаш бойдустуң болуушкуннары/ Природные объекты и явления		*соок* ‘холодный ’, *кадыр* ‘крутой’,
Чүүл/Предмет	физические свойства	*быжыг* ‘крепкий’, *суук* ‘жидкий’, *улуг* ‘большой’.
	цвет	*кызыл* ‘красный’, *сарыг* ‘желтый’.
Демдек/Оценка	положительные	*эки* ‘хороший’, *чараш* ‘красивый’.
	отрицательные	*багай* ‘плохой’, *чүдек* ‘некрасивый’.

Лексико-семантические пометы:
Относительные

Кижи/Человек ***Дириг амытан***/ Животное	***демдек***/Признак	*аъттыг* ‘имеющий лошадь’, *малдыг* ‘имеющий скот’.

Чүүл/Предмет, ***Бойдус объектилери болгаш бойдустуң болуушкуннары***/ Природные объекты и явления	***үе***/Время	*кышкы* ‘зимний ’, *чайгы* ‘летний’
Чүүл/Предмет	***туруш***/Место	*тайгадагы* ‘находящийся в тайге’.

Глаголы

Лексико-семантические блоки:

- Действие
- Деятельность
- Бытие
- Качество
- Состояние
- Отношение.

Лексико-семантические пометы:

Кылдыныг/ Действие	***шимчээш-кин***/ движение	***Кижи***/Человек	*халы-* ‘бегать’, *дедирлен-* ‘двигаться назад’, *күзегле-* ‘перекочевывать на осеннее стойбище’, *чайлагла-* ‘перекочевывать на летнее стойбище’
		Дириг амытан/ Животное	*даалыкта-* ‘скакать галопом’, *кылыйт-* ‘очень быстро лететь (о птице)’, *сояста-* ‘ползти, ползать (о пресмыкающемся)’.

		Бойдус объектилери болгаш бойдустуң болуушкуннары/ Природные объекты и явления	*ак-* ‘течь, струиться вниз’, *бысканна-* ‘идти о мелком снеге, порошить’, *дүргектел-* ‘клубиться’.
		Чүүл/Предмет	*ужук-* ‘лететь’, *эсте-* ‘лететь по ветру’.
	объектиже угланган кылдыныг/ физическое воздействие на объект	***Кижи***/Человек	*хак-* ‘бить, ударять, стучать, колотить; *тапта-* топтать, растаптывать, *суйба-* ‘гладить ’.
		Дириг амытан/ Животное	*ызыр-* ‘кусать’, *чазарла-* ‘растерзать’.
	физиологтуг кылдыныг/ физиологическое действие	***Кижи***/Человек	*дайна-* ‘жевать, пережевывать’, *дерит-* ‘потеть’, *оптук-* ‘поперхнуться’.
		Дириг амытан/ Животное	*бышкыр-* ‘чихать (о козе и овце)’, *төрү-* ‘родить (о животных)’.
	ыыт-даап, ун үндүрери/ звучание	***Кижи***/Человек, ***Дириг амытан***/ Животное, ***Чүүл***/ Предмет, ***Бойдус объектилери болгаш бойдустуң болуушкуннары***/ Природные объекты и явления	*диңмире-* ‘греметь, грохотать (о громе) ’, *кагжыра-* ‘шуршать, скрипеть, хрустеть’, *койтула-* ‘булькать’, *коола-* ‘выть, завывать’.

Ажыл чорудулгазы/ Деятельность	***кылып бүдүрери***/ созидательная	***Кижи***/Человек	*сыры-* ‘стегать, простёгивать, прошивать’ *чон-*‘тесать’, *шутку-* ‘отливать, изготовлять литьём’.
	угаан ажылы/ интеллектуальная		*бода-* думать, раздумывать, мыслить; предполагать, *ожаа-* ‘замечать; отмечать, принимать во внимание’.
	чугаа чорудулгазы/ речевая		*ааза-* ‘обещать, сулить’, *де-* ‘говорить, сказать’, *нүгүлде-* ‘клеветать’.
	ниитилел ажыл чорудулгазы/ социальная		*башкыла-* ‘преподавать, учительствовать’, *кадарчыла-* ‘быть пастухом, чабаном ’.
Бар болуру/ Бытие	***тыптып келири***/ начало бытия	***Кижи***/Человек	*төрүттүн-* ‘родиться’, *боттан-* ‘зачинаться, появляться’.
		Дириг амытан/ Животное	*төрү-* ‘родить (о животном)’.
		Бойдус объектилери болгаш бойдустуң болуушкуннары/ Природные объекты и явления	*кел-* ‘наступать, наставать’, *эгеле-* ‘начинаться’, *дүш-* ‘наступать, наставать’.
		Чүүл/Предмет	*тывыл-* ‘возникать, зарождаться, появляться’.
	бар болуру/ существование	Кижи/Человек	*амыдыра-* ‘жить’, *сыңмарлаш-* ‘тесниться, ютиться’, *чайлагла-* ‘жить на летней стоянке’, *чуртта-* ‘жить’.

		Дириг амытан/ Животное	*хоргада-* ‘ютиться’, *таварыш-* ‘встречаться, существовать’).
		Бойдус объектилери болгаш бойдустуң болуушкуннары/ Природные объекты и явления	*дургел-* ‘висеть о тучах’, *чаттыл-* ‘тянуться’, *өс-* ‘расти’.
		Чүүл/Предмет	*чору-* ‘быть, бывать’, *бол-* ‘быть, бывать’.
	Чиде бээри/ прекращение бытия	***Кижи***/Человек	*мөчү-* ‘скончаться’, *өл-* ‘умирать, погибать’, *дүш-* ‘погибать в воде’, *кал-* ‘умирать’
		Дириг амытан/ Животное	*өл-* ‘дохнуть, подыхать’, *кырыл-* ‘издыхать’
		Бойдус объектилери болгаш бойдустуң болуушкуннары/ Природные объекты и явления	*бустал-* ‘испаряться’, *курга-* ‘высыхать, сохнуть’.
		Чүүл/Предмет	*балал-* ‘исчезать, стираться, выводиться, изглаживаться’, *буура-* ‘разваливаться’, *сен-* ‘рассасываться’, *чайла-* ‘миновать’.
Шынар/ Качество	***демдек***/ признак	***Кижи***/Человек	*дадык-* ‘крепнуть, закаляться’, *кыры-* ‘стареть’, *ар-* ‘худеть’, *ижик-* ‘вживаться, привыкать’.

		Дириг амытан/ Животное	*быжык*- ‘крепнуть’.
		Бойдус объектилери болгаш бойдустуң болуушкуннары/ Природные объекты и явления	*соо*- ‘остывать, охлаждаться, холодеть’, *дошта*л- ‘леденеть’, *ири*- ‘гнить, разлагаться’.
		Чүүл/Предмет	*дашта*л- ‘каменеть’, *сандара*- ‘ветшать’, *самдара*- ‘оборваться, растрепаться’.
	өң/цвет	***Кижи***/Человек	*додук*- ‘загореть’, *долбаннал*- ‘румяниться’.
		Дириг амытан/ Животное	*карар*- ‘буреть, чернеть’.
		Бойдус объектилери болгаш бойдустуң болуушкуннары/ Природные объекты и явления	*ногаарар*- ‘зеленеть’, *саргар*- ‘желтеть’.
		Чүүл/Предмет	*агар*- ‘белеть, седеть’, *оң*- ‘блекнуть’, *мөнгүннел*- ‘серебриться’.
	Сагыш-сеткилдиң илерээри/ эмоция	***Кижи***/Человек	*дүвүре*- ‘беспокоиться, волноваться, тревожиться’, *кайга*- ‘удивляться’, *кудара*- ‘грустить, тосковать’

		Бойдус объектилери болгаш бойдустуң болуушкуннары/ Природные объекты и явления	
Байдал/ Состояние	***физиологтуг байдал***/ физиология	***Кижи***/Человек	*аары-* ‘болеть’, *шыла-* ‘уставать’, *ашта-* ‘голодать’, *көжү-* ‘застывать’, *божу-* ‘родить’
		Дириг амытан/ Животное	*чуда-* ‘тощать (о животных)’
Хамаарылга/ Отношение	***кижилер аразында хамаарылгалар***/ социальные отношения	***Кижи***/Человек	*өгленир* ‘жениться, выйти замуж’, *харылзажыр* ‘взаимодействовать’, *деткиир* ‘поддрежиавать’.

Наречие

Лексико-семантические пометы:

Кижи/Человек, ***Дириг амытан***/ Животное, ***Чүүл***/Предмет, ***Бойдус объектилери болгаш бойдустуң болуушкуннары***/ Природные объекты и явления	***туруш***/место	*даштын* ‘на улице’, *дүгде* ‘там’
	угланыышкын/ направление	*өскээр* ‘в другую сторону’, *аткаар* ‘назад’, *даштыыртан* ‘снаружи’
	үе/время	*ам* ‘сейчас’, *даарта* ‘завтра’, *кежээликтей* ‘вечерком’, *хүнзедир* ‘целый день’
	дүргени/скорость	*дүрген* ‘быстро’, *таваар* ‘не спеша’, *үр* ‘долго’
	чадазы/степень	*хөлчок* ‘очень’, *аажок* ‘очень’, *шала* ‘немного’, *дыңзыдыр* ‘туго’, *кошкак* ‘слабо’, тоттур ‘досыта’
Кижи/Человек	***сорулга***/цель	*сагыштыы-биле* ‘специально’, *хей-ле* ‘зря’

Местоимение

Лексико-семантические пометы:

Кижи/Человек, ***Дириг амытан***/ Животное, ***Чүүл***/Предмет, ***Бойдус объектилери болгаш бойдустуң болуушкуннары***/ Природные объекты и явления	***арын***/лицо	*мен* ‘я’, *кым* ‘кто’, *шупту* ‘все’, кым-бир ‘кто-то’
	демдек/признак	*ындыг* ‘такой’, *кандыг* ‘какой’
	үе/время	*ынчан* ‘тогда’, *кажан* ‘когда’, *кажан-бир* ‘когда-нибудь’
	угланыышкын/ направление	*ынаар* ‘туда’, *кайнаар* ‘куда’
	туруш/место	*ында* ‘там’, *кайда* ‘где’
	сан/число	*ынча* ‘столько’, *чеже* ‘сколько’, *элээн* ‘немного’
	чада/степень, ***хемчээл***/мера	*ол хире* ‘столько’, *кайы хире* ‘насколько’
	чылдагаан/ причина	*чүге* ‘почему’

Лексико-семантические пометы даются на тувинском и русском языках, поскольку в дальнейшем программа автоматизированной семантической обработки текстов в качестве подкорпуса ЭКТЯ будет адресована не только для создателей учебника, учителей, но и для учащихся.

Имея такого рода семантическую разметку, можно быстро подготовить учебно-методические комплексы не только по тувинскому языку, но и по литературе, а также подготовить электронно-образовательные ресурсы для школьников.

ЛИТЕРАТУРА

1. Бавуу-Сюрюн М. В., Далаа С. М. Морфемно-орфографический словарь тувинского языка. Электронный ресурс.

2. Хертек А.Б., Ооржак Б.Ч., Ондар В.С., Салчак А.Я., Соян А.М. Словарь тематических групп глагольной лексики тувинского языка. Электронный ресурс.

LINGUISTIC ANNOTATION OF GRAMMATICAL CATEGORIES OF SAKHA LANGUAGE (ON EXAMPLE OF NOUN)

Gavril Torotoev, Alina Torotoeva

M.K. Ammosov North-Eastern Federal University
Yakutsk, Sakha, Russia

This paper shows the work to create instruments for linguistic annotation of grammatical categories of Sakha. It describes the basic inflectional characteristics of nouns of Yakut language (numbers, personal endings, possessive endings, cases), which are based on Leipzig Glossing Rules.

Лингвистическое аннотирование грамматических категорий языка является актуальной проблемой современной компьютерной лингвистики. В последнее время в связи с интенсивным развитием компьютерных технологий назрела необходимость в разработке системы грамматической разметки для автоматического анализа текстов, хранящихся в электронных корпусах тюркских языков. В целях повышения эффективности работы языковедов при проведении сравнительно-сопоставительных исследований и для получения объективных языковых данных также необходима унификация системы условных сокращений.

Для морфологического аннотирования лексико-грамматических разрядов языка саха нами используется система тэгов, базирующаяся на Лейпцигских правилах глоссирования (табл. №1).

Таблица №1

Части речи в якутском языке

Сокращения	Расшифровка сокращений	Название категории
N	Noun	Имя существительное
POSS	Possessive	Имя притяжательное
PRO	Pronoun	Местоимение
NUM	Numeral	Имя числительное
ADJ	Adjective	Имя прилагательное
V	Verb	Глагол
PCP	Participle	Причастие

CONV	Converb	Деепричастие
ADV	Adverb	Наречие
MOD	Modal word	Модальное слово
INTJ	Interjection	Междометие
CONJ	Conjunction	Союз
PART	Particle	Частица
POST	Postposition	Послелог
IMIT	Imitative word	Звукоподражательное слово

Якутский язык относится к числу тюркских языков, но в данной языковой группе он стоит обособленно в силу определенных культурно-исторических факторов. Академик О.Н. Бетлингк в своем классическом труде «О языке якутов» (1851) предполагает, что «якуты первые отделились от членов турецко-татарской семьи, еще не разделившихся в отношении языка». [1]. Он считает, что язык саха «является древнейшим из всех нынешних тюркских языков». [2]. Его научная гипотеза подтверждается объективными данными современных тюркологических исследований. Якутский язык в своем развитии испытал некоторое влияние монгольского и тунгусо-маньчжурского языков. Те или иные исторические обстоятельства внесли определенные корректировки в языке народа саха, обогатив тем самым систему грамматических категорий якутского языка.

В таблице №2 в качестве дидактической единицы продемонстрированы универсальные и специфические частеречные характеристики английского, русского и якутского языков. Ознакомившись с содержанием таблицы, обучающиеся приходят к выводу, что унифицированные тэги образованы из английских терминов. Звездочкой обозначены те категории, грамматические значения которых в русском или якутском языках совпадают лишь частично. Система условных сокращений, основанная на якутских терминах, была разработана нами для внутреннего пользования, иначе говоря, студент использует данные тэги тогда, когда он производит лингвистический анализ на своем родном языке.[3]. Подобные сопоставительные таблицы помогают студентам самостоятельно находить языковые универсалии и параллели в разноструктурных языках.

Таблица №2

Части речи в английском, русском и якутском языках и их условные обозначения

Английские термины	Русские термины	Якутские термины	Унифицированные тэги	Тэги собственной разработки
Nouns	Имена существительные	Аат тыл	N	А
Possessive Case of Nouns*	Имена притяжательные	Тардыылаах аат тыл	POSS	А//
Pronouns	Местоимения	Солбуйар аат	PRO	СА
Numerals	Имена числительные	Ахсаан аат	NUM	АА
Adjectives	Имена прилагательные	Даҕаанын аат	ADJ	Д
Verbs	Глагол	Туохтуур	V	Т
Participles*	Причастия	Аат туохтуур	PCP	АТ
–	Деепричастия	Сыһыат туохтуур	CONV	СТ
Adverbs	Наречия	Сыһыат	ADV	С
Modal words*	Модальные слова*	Сыһыан тыл	MOD	СыһТ
Interjections	Междометия	Саҥа аллайыы	INTJ	САл
Conjunctions	Союзы	Ситим тыл	CONJ	Сит
Particles*	Частицы	Эбиискэ	PART	Э
Prepositions	Предлоги	–	PREP	–
Postpositions*	–	Дьөһүөл	POST	Дь
–	Звукоподражательные слова	Тыаһы үтүктэр тыл	IMIT	ТУТ
Articles	–	–	ART	–

*Частичное соответствие

В данной статье словообразовательные характеристики имени существительного не будут рассмотрены обстоятельно, так как словообразование для автоматической обработки текста не является первостепенной. К нему обычно обращаются после описания словоизменительных характеристик языка для того, чтобы составить полную базу аффиксов для дальнейших этимологических исследований. В образовании существительных участвует много продуктивных и малопродуктивных аффиксов: –һыт (-сыт, -чыт, -дьыт, -ньыт), -был (-бил, -бул, -бүл), -ык (-ик, -ук, -үк), -ах (-эх, -ох, -өх), -лаҥ (-лэҥ, -лоҥ, -лөҥ), -лта (-лтэ, -лто, -лтө) и др. Но в рамках данной статьи словообразовательные возможности имени существительного остаются вне зоны нашего исследования. В качестве образца описания словообразовательных аффиксов обратимся исключительно к 3 часто употребляемым морфемам, образующим отглагольные имена существительные (табл. №3).

Таблица №3

Деривация

Сокращения	Расшифровка сокращений	Название категории	Алломорфы (12)	Морфемы (3)
AN	Agens noun	имя деятеля на –ааччы (присоединяется к любой глагольной основе)	-ааччы/ -ээччи/ -ооччу/-өөччү	-ааччы
VN_1	Verbal noun	имя действия на –ыы (присоединяется к глагольным основам с конечным согласным)	-ыы/-ии/-уу/ -үү	-ыы
VN_2	Verbal noun	имя действия на –ааһын (присоединяется к глагольным основам с конечным долгим гласным)	-ааһын/ -ээһин/ -оонун/-өөһүн	-ааһын

Как правило, особое внимание ученых, занимающихся проблемами компьютерной лингвистики, сконцентрировано на словоизменительную морфологию. Далее детально рассмотрим основные словоизменительные характеристики имени существительного якутского языка: число (см. табл. №4), падеж (см. табл. №5 и №7), посессивность (см. табл. №6) и лицо (см. табл. №6 и №8).

Таблица №4

Категория числа

Сокращения	Расшифровка сокращений	Название категории	Алломорфы	Морфемы
SG	singular	единственное число	–	
PL	plural	множественное число	-лар/-лор/-лэр/-лөр -нар/-нор/-нэр/-нөр -дар/-дор/-дэр/-дөр -тар/-тор/-тэр/-төр	-лар

В якутском языке аффикс множественного числа *–лар* представлена шестнадцатью (16) формальными показателями. При выборе оптимального алломорфа решающую роль играет закон сингармонизма в якутском языке. Фонетическая сочетаемость морфем также зависит от правил ассимиляции (например, прогрессивная, регрессивная, прогрессивно-регрессивная ассимиляция согласных) и аккомодации. Таким образом, согласно закону сингармонизма, правилам ассимиляции и аккомодации, разрабатываются правила сандхи, показывающие изменения звуков на морфемных швах.

В нашей работе, в частности, в идентификации грамматических категорий с соответствующими тэгами очень помог научный труд О.Н. Бетлингка «О языке якутов», который был издан на немецком языке в далеком 1851 г. Академик предполагает, что в якутском языке имеется 10 падежей: Casus Indefinitus, Accusativus indefinites, Dativ, Accusativus definitus, Ablativ, Lokativ, Instrumental, Casus adverbialis, Comitativ, Casus Comparativus [4]. Все падежи якутского языка именуются согласно принятому тогда в языкознании терминами, имеющими латинские корни. И поэтому нет существенных расхождений между терминами, используемыми

О.Н. Бетлингком и современными унифицированными тэгами. В современном якутском языке выделяют 8 падежей, не вошли в падежную парадигму Lokativ (местный) и Casus adverbialis (творительный).

В якутском языке различаются два типа склонения: простое и притяжательное (см. табл. №5 и №7). В простом склонении все морфемы имеют по четыре алломорфа, например: -ҥа (-ҥа/-ҥо/-ҥэ/-ҥө).

Таблица №5

Простое склонение

Унифицированные сокращения	Название категории в работе О.Н. Бетлингка «О языке якутов» (1851)	Название категории в «Грамматике современного якутского литературного языка» (1982)	Алломорфы (92)	Морфемы (10)
NOM	Неопределённый падеж (Casus Indefinitus)	Основной падеж (Төрүт түһүк)	–	–
ACC_INDF	Винительный неопределённый падеж (Accusativus indefinitus)	Частный падеж (Араарыы түһүк)	-та/-то /-тэ/-тө -ла/-ло /-лэ/-лө -на/-но /-нэ/-нө -да/-до /-дэ/-дө	-та
DAT	Дательный падеж (Dativ)	Дательный падеж (Сыһыарыы түһүк)	-ҕа/-ҕо/-ҕэ/-ҕө -ха/-хо/-хэ/-хө -га/-го/-гэ/-гө -ҥа/-ҥо/-ҥэ/-ҥө -ка/-ко-/-кэ/-кө	-ҕа
ACC_DEF	Винительный определённый падеж (Accusativus definitus)	Винительный падеж (Туохтуу түһүк)	-ы/-и/-у/-ү -ны/-ни/-ну/-нү	-ы -ны
ABL	Исходный падеж (Ablativ)	Исходный падеж (Таһаарыы түһүк)	-тан/ -тон/-тэн/-төн -ттан/ -ттон/-ттэн/ -ттөн	-тан -ттан

INS	Орудный падеж (Instrumental)	Орудный падеж (Туттуу түһүк)	-нан/-нон/-нэн/-нөн -ынан/-унан/-инэн/-үнэн	-нан -ынан
COM	Совместный падеж (Comitativ)	Совместный падеж (Холбуу түһүк)	-лыын/-лиин/-луун/-лүүн -ныын/-ниин/-нуун/-нүүн -тыын/-тиин/-туун/-түүн -дыын/-диин/-дуун/-дүүн	-лыын
COMP	Сравнительный падеж (Casus comparativus)	Сравнительный падеж (Тэҥнии түһүк)	-тааҕар/-тооҕор/-тээҕэр/-төөҕөр -наaҕар/-нооҕор/-нээҕэр/-нөөҕөр -даaҕар/-дооҕор/-дээҕэр/-дөөҕөр -лаaҕар/-лооҕор/-лээҕэр/-лөөҕөр	-тааҕар

Категория принадлежности в языке саха представлена 7 аффиксами с их 59 фонетическими вариантами (табл. №6). Данные формы обладают высокой употребительностью. «Они выражают разнородный круг логических отношений и связей между предметами, нередко далеких от понятия принадлежности или обладания». [5].

Таблица №6

Категория принадлежности

Сокращения	Название категории	Алломорфы	Морфемы
POSS_1SG	Принадлежность 1 лицу единственного числа	-м -ым/-им/-ум/-үм	-м
POSS_2SG	Принадлежность 2 лицу единственного числа	-ҥ -ыҥ/-иҥ/-уҥ/-үҥ	-ҥ
POSS_3SG	Принадлежность 3 лицу единственного числа	-а/-о/-э/-ө -та/-то/-тэ/-тө	-а -та

POSS_1PL	Принадлежность 1 лицу множественного числа	-быт/-бит/-бут/-бүт -пыт/-пит/-пут/-пүт -мыт/-мит/-мут/-мүт	-быт
POSS_2PL	Принадлежность 2 лицу множественного числа	-ҕыт/-ҕит/-ҕут/-ҕүт -хыт/-хит/-хут/-хүт -кыт/-кит/-кут/-күт -гыт/-гит/-гут/-гүт -ҥыт/-ҥит/-ҥут/-ҥүт	-ҕыт
POSS_3PL	Принадлежность 3 лицу множественного числа	-лара/-лоро/-лэрэ/-лөрө -нара/-норо/-нэрэ/-нөрө -дара/-доро/-дэрэ/-дөрө -тара/-торо/-тэрэ/-төрө	-лара

В якутском языке в простом и притяжательном склонениях используются 184 алломорфа. Данный показатель свидетельствует об огромном функциональном потенциале имени существительного как особого лексико-грамматического разряда слов.

Таблица №7

Притяжательное склонение

Унифицированные сокращения	Название категории в работе О.Н. Бетлингка «О языке якутов» (1851)	Название категории в «Грамматике современного якутского литературного языка» (1982)	Алломорфы (92)	Морфемы (15)
NOM	Неопределенный падеж (Casus Indefinitus)	Основной падеж (Төрүт түһүк)	<u>1-е лицо ед.числа:</u> -м -ым/-им/-ум/-үм <u>2-е лицо ед.числа:</u> -ҥ -ыҥ/-иҥ/-уҥ/-үҥ <u>3-е лицо ед.числа:</u> -а/-о/-э/-ө -та/-то/-тэ/-төрө <u>1-е лицо мн.числа:</u> -быт/-бит/-бут/-бүт -пыт/-пит/-пут/-пүт -мыт/-мит/-мут/-мүт	-м -ым -ҥ -ыҥ -а -та -быт

			2-е лицо мн.числа: -ҕыт/-ҕит/-ҕут/-ҕүт -хыт/-хит/-хут/-хүт -кыт/-кит/-кут/-күт -гыт/-гит/-гут/-гүт -ҥыт/-ҥит/-ҥут/-ҥүт 3-е лицо мн.числа: -лара/-лоро/-лэрэ/-лөрө -нара/-норо/-нэрэ/-нөрө -дара/-доро/-дэрэ/-дөрө -тара/-торо/-тэрэ/-төрө	-ҕыт -лара
ACC_INDF	Винительный неопределенный падеж (Accusativus indefinitus)	Частный падеж (Араарыы түһүк)	-ына/-инэ/-уна/-үнэ	-ына
DAT	Дательный падеж (Dativ)	Дательный падеж (Сыһыарыы түһүк)	-ар/-ор/-эр/-өр -гар/-гэр	-ар -гар
ACC_DEF	Винительный определенный падеж (Accusativus definitus)	Винительный падеж (Туохтуу түһүк)	-н -ын/-ин/-ун/-үн	-н -ын
ABL	Исходный падеж (Ablativ)	Исходный падеж (Таһаарыы түһүк)	-ттан/-ттэн	-ттан
INS	Орудный падеж (Instrumental)	Орудный падеж (Туттуу түһүк)	-нан /-нэн	-нан
COM	Совместный падеж (Comitativ)	Совместный падеж (Холбуу түһүк)	-ныын/-нуун/-наан /-нээн (-ниин)	-ныын
COMP	Сравнительный падеж (Casus comparativus)	Сравнительный падеж (Тэҥнии түһүк)	-нааҕар/-нээҕэр	-нааҕар

Имена существительные якутского языка в предложении могут выступать в роли сказуемого. В зависимости от контекста в таких случаях к основам слов прибавляются определенные аффиксы сказуемости за исключением 3 лица единственного числа.

Таблица №8

Аффиксы сказуемости имен существительных

Сокращения	Название категории	Алломорфы	Морфемы
1SG	1 лицо единственного числа	-бын/-бин/-бун/-бүн -мын/-мин/-мун/-мүн -пын/-пин/-пун/-пүн	-бын
2SG	2 лицо единственного числа	-ҕын/-ҕин/-ҕун/-ҕүн -хын/-хин/-хун/-хүн -кын/- кин/-кун/-күн -гын/-гин/-гун/-гүн -ҥын/-ҥин/-ҥун/-ҥүн	-ҕын
3SG	3 лицо единственного числа	–	–
1PL	1 лицо множественного числа	после аффикса –лар: -быт/-бит/-бут	-быт
2PL	2 лицо множественного числа	после аффикса –лар: -гыт/-гит/-гут	-гыт
3PL	3 лицо множественного числа	-лар/-лор/-лэр/-лөр -нар/-нор/-нэр/-нөр -дар/-дор/-дэр/-дөр -тар/-тор/-тэр/-төр	-лар

В статье в общей сложности охвачено и интерпретировано грамматическое значение 268 аффиксов, в том числе словообразовательных -3, словоизменительных – 265.

А теперь, используя тэги, обозначающие те или иные грамматические категории, произведем анализ имени существительного. Для репрезентативности практического материала сначала обратимся к морфемному анализу, затем к морфологическому аннотированию.

Пример 1. *Оҕоҕун* 'твоего ребенка'

а) Морфемный анализ: оҕо+ҕ+ун. Расшифровка: имя существительное + принадлежность 2 лицу единственного числа (-ҥ) + винительный падеж.

б) Морфологическое аннотирование: N- POSS_2SG -ACC_DEF

Пример 2. *Суруйааччыларбытыгар* 'нашим писателям'

а) Морфемный анализ: суруйааччы+лар+быт +(ы)гар. Расшифровка: имя деятеля на -ааччы + множественное число + принадлежность 1 лицу множественного числа + дательный падеж.

б) Морфологическое аннотирование: AN-PL-POSS_1PL-DAT

Пример 3. *Аттаргытынааҕар* (түргэн) '(быстрее) ваших коней'

а) Морфемный анализ: ат+тар+гыт+(ы)нааҕар. Расшифровка: имя существительное + множественное число (-лар) + принадлежность 2 лицу множественного числа (-ҕыт) + сравнительный падеж.

б) Морфологическое аннотирование: N-PL-POSS_2PL- COMP

Таким образом, в данной статье систематизирована и апробирована система тэгов, отражающая словоизменительный потенциал имени существительного. Для того, чтобы компьютер мог автоматически проанализировать тексты любой сложности, представленном в электронном корпусе якутского языка, необходимо описать унифицированными тэгами все грамматические категории языка саха. При решении данной проблемы станет возможным создание новых компьютерных программ, таких как онлайн-переводчик, автоматический анализатор текстов и др.

СПИСОК ЛИТЕРАТУРЫ

1. Бетлингк О.Н. О языке якутов / Пер. с нем. Рассадин В.И. – Новосибирск: Наука. Сиб. Отд-ние, 1990. – С. 44.

2. Грамматика современного якутского литературного языка. Фонетика и морфология. – Москва: Наука, 1982. – С. 10.

3. Торотоев Г.Г. Метод моделирования в исследовании стихообразующего каркаса олонхо // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2014. – Казань: Изд-во «Фэн» Академии наук РТ, 2014. – С. 243–247.

4. Бетлингк О.Н. О языке якутов / Пер. с нем. Рассадин В.И. – Новосибирск: Наука. Сиб. Отд-ние, 1990. – С. 278–285.

5. Грамматика современного якутского литературного языка. Фонетика и морфология. – Москва: Наука, 1982. – С. 129.

SECTION 5 | **SEMANTICS AND GRAMMAR IN TURKIC LANGUAGES**

PREDICATE-ARGUMENT RELATIONS IN KOREAN SENTENCES WITH PARTICIPLE PHRASES

Evgeniya Brechalova

Russian State University for the Humanities, Institute for Oriental and Classical Studies, Miusskaya sqr. 6-1, Moscow, 125993, Russia
evbrechalova@gmail.com

The paper concerns rules for establishing predicate-argument relations in Korean complex sentences with adnominalizations, i.e. with clauses whose main predicate is put in an adnominalized, or participle form (Part) to modify the following noun (Host), syntactic head of the Part. In a sentence with participles, the rules evaluate whether a noun (Act) marked by subject or object particle and located somewhere to the left of the Part is an actant of the Part or an actant of the other predicate (Q) located to the right of the Part. Before applying the rules to a sentence, it should be divided into so-called platforms. A platform ends with a noun with topic particle or a verb/adjective form such as converb, nominalization, finite form. It is important that the analytical verb forms built on the basis of participles and auxiliary nouns are counted as quasi-converbs. Thus, the proposed rules apply only to simple participles, not to quasi-converbs, and only to those subject/object marked nouns that are located in the same platform with the participle. The rules have no need to employ any type of lexical dictionary. The paper claims that an Act is an actant of a Part, if one of the following conditions is fulfilled: (1) the Act and the Part are both located within a topical platform, (2) the Host is one of a class of predicative nouns (e.g. kes a fact, il a matter; an affair; kwangkyeng a sight), (3) the Part is located in finite platform while there is also a topical platform in the sentence, (4) particles marking the Host and the Act are the same. Since Korean is one of the Altai languages, it is typologically very close to Turkic languages. So the paper may be of interest to those involved in the automatic analysis of Turkic languages.

1. Введение

В настоящем докладе обсуждается частная задача установления актантно-предикатных связей в корейском сложном предложении с причастными оборотами, т.е. придаточными предложениями, сказуемое которых имеет форму причастия. При этом все придаточное предложение играет роль определения к некоторому последующему имени. При анализе сложного предложения, в котором присутствуют причастные формы, возникает вопрос о том, где находится левая граница причастного оборота, поскольку в причастный оборот или adnominalization, как подтверждается грамматикой (Мартин, 1992; 11.9) может быть превращено предложение какой угодно длины. В частности, с точки зрения задачи об установлении актантно-предикатных связей, необходимо решить про имена с актантными показателями, находящиеся в препозиции по отношению к причастию, являются ли они актантами причастия – т.е. входят в причастный оборот, или являются актантами некоторого последующего, принимающего причастный оборот, предикатива. Среди актантов важнейшую роль играют d1, маркируемый номинативом, и d2, маркируемый аккузативом. Именно они попадают в область внимания в настоящей работе.

Естественно, что задача о “содержимом” причастного оборота связана с проблемой построения поверхностно-синтаксического представления корейского предложения. Решение этой задачи значительно упрощает соответствие между синтаксическим и семантическим представлением предложения, процедура и правила установления которого описана в (Brechalova, 2010).

Настоящая работа – это продолжение работы над синтаксическим анализатором для корейского как языка алтайского типа. Поскольку типологически тюркские языки и корейский язык очень близки, то данная работа может оказаться небезынтересной для автоматического анализа тюркского синтаксиса.

2. Исходная разметка текста: платформы

В настоящей работе используется специальная разметка корейского текста. А именно, корейское предложение делится на некоторые линейные отрезки, так называемые платформы. Граница платформы проходит либо после имени с показателем топика,

либо после предикатива в одной из форм финитного или срединного сказуемого. Деление на платформы отражает вид разрезки предложения на группу топика и цепочку простых предложений в составе сложного. Платформы бывают нескольких типов, в зависимости от граммматической характеристики той словоформы, на которую заканчивается платформа: FIN (финитная), TOP (топикальная), GER (деепричастная), NMZ (номинализации), PART-QUOT (сказуемое придаточного косвенной речи в определительной позиции), GER-QUOT (сказуемое придаточного косвенной речи в позиции обстоятельства).

В таблице (1) приведено предложение, в котором деление на платформы указано квадратными скобками, а тип платформы – верхними индексами:

Таблица 1

Образец исходной разметки текста

[oykwuk=eyse	mantul.eci=n	mwulken=i	swuiptoy=e]$_{GER}$	[hankwuk=eyse	phanmaytoy=nun
N=Loc	V=Part	N=Nom	V=Ger	N=Loc	V=Part
заграница	быть сделанным	вещь	ввозить	Корея	продавать
kyengwu,]$_{GER}$	[kakyek=i	wenka=pota	noph=un	kyengwu=ka	manh=supnita]$_{FIN}$
N=Ø	N=Nom	N=Compar	Adj=Part	N=Nom	Adj=Fin. Decl
случай	цена	оригинальная цена	высокий	случай	много
Если в Корее продают товар, произведенный за границей, то во многих случаях его цена (в Корее) оказывается выше, чем оригинальная цена в стране-производителе.					

Для решения поставленной задачи важны два обстоятельства. Во-первых, некоторые сказуемые по форме представляют собой аналитические конструкции вида «причастие + служебное имя» или «причастие + служебное имя + служебный предикатив». Подобные сказуемые приравниваются к деепричастным или финитным формам. В предложении (1) таково срединное сказуемое phanmaytoy=nun kyengwu *если продают*. Во-вторых, важно, что после причастий, не входящих в состав аналитических конструк-

ций, не проводится граница платформы. Подобные простые причастия всегда входят внутрь какой-либо платформы. В предложении (1) простые причастия mantul.eci=n *быть сделанным* и noph=un *высокий* входят внутрь Ger и Fin платформ соответственно.

Именно актантно-предикатные связи простых причастий и строение платформ, в которых они встретились, являются объектом исследования в настоящей работе.

3. Поверхностно-синтаксическое представление и актанты причастий

Результаты решения задачи об актантах причастий используются в поверхностно-синтаксическом представлении особого сорта. Оно имеет вид леса деревьев, где каждое отдельное дерево отображает одну платформу. При построении данного представления используются правила допустимых непосредственных составляющих; эти правила, учитывающие левый контекст текущей словоформы, сформулированы в (Бречалова, 2008). Ниже иллюстрируется соответствие между предложением и его поверхностно-синтаксическим представлением: предложению (см. Табл.2), содержащему 4 платформы, отвечает лес из 4 деревьев (см. Рис.1), причем в данном примере одно из них представлено только корневой вершиной.

Таблица 2

[pam	sai			phokphwung=i	memchwu=ess=ten=ci]$_{NMZ}$
N= Ø	N= Ø			N=Nom	V=Past=Nmz
ночь	в течение			ураган	кончиться
[otwumak=ey=nun	enusay	hayspich=i	katuk.ha=ess=ko]^{GER}		
N=Dat=Top	N=Ø	N=Nom	V=Past=Ger		
хижина	какой-то момент	солнечный свет	наполниться		
[haykulitu=nun]$_{TOP}$	[phwuk	kkeci=n	sopha=eyse	(camtul=e+iss=ess=ta)]^{FIN}	
N=Top	N=Ø	V=Part	N=Abl	V=Progress.Past.Fin.Decl	
Хагрид	полностью	сломаться	софа	спать	
Хижина была залита светом, ураган кончился, Хагрид спал на сломанной софе.					

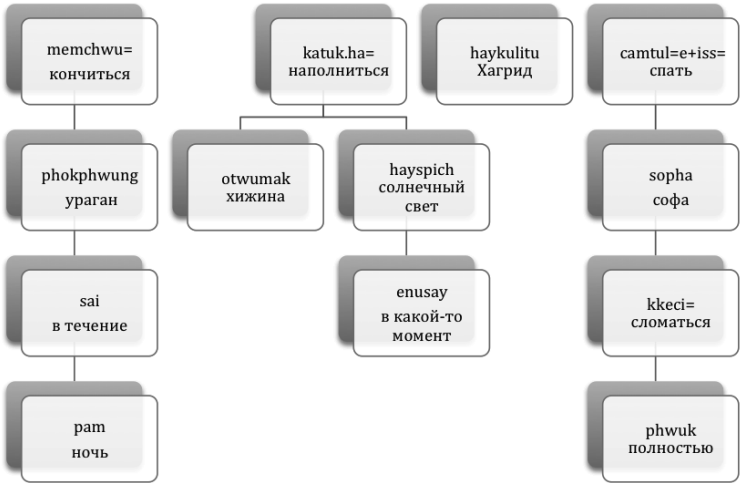


Рис. 1.

Ситуация с определением хозяина для имени с показателем номинатива или аккузатива оказывается неоднозначной, если внутри платформы встретилась конфигурация вида

[Act Part Host],

где Act – это именная словоформа с показателем номинатива или аккузатива, расположенная между началом платформы и причастием Part, а Host – это синтаксический хозяин причастия, определением к которому служит причастие.

Во-первых, словоформа Act может быть актантом причастия Part. Образец этой ситуации представлен ниже одним предложением (см. Табл.3) и его поверхностно-синтаксической структурой (см. Рис.2).

Таблица 3

Act – актант для Part: платформы

[ku=nun]^{TOP}	[tto	mwusun	ttus	i=n=ci=to]^{NMZ}	[molu=keyss= ciman]^{GER}
N=Top	N= Ø	N= Ø	N= Ø	Copul=Nmz	V=Fut=Ger
он	также	какой	смысл	быть	не знать

[**casin=i**	‘mekul=lo’	**pwulli=ess=ten**	kes=ey	tayha=e]^{GER}	[sayngkak.ha=ess=ta]^{FIN}
N=Nom **ACT**	N=Ins	V=Past=Part.Retr **PART**	N=Dat	V=Ger	V=Past=Fin.Decl
сам	магл	быть названным	факт	относительно	думать
Мало того, его назвали каким-то маглом. Что бы там ни означало это слово, мистер Дурсль был потрясен.					

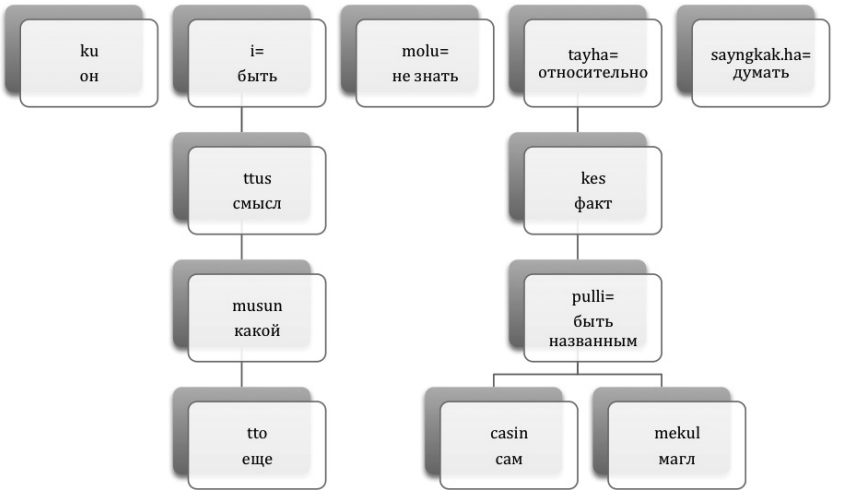


Рис. 2. Act – актант для Part: поверхностно-синтаксическое представление

Во-вторых, словоформа Act может быть актантом Q, главного предикатива платформы, после которого проходит ее граница. Образец этой ситуации представлен ниже одним предложением (см. Табл.4) и его поверхностно-синтаксической структурой (см. Рис.3).

Таблица 4

Act – актант для Q: платформы

[ku=tul=un]^{TOP}	[**twutulli=ka**	kuleh=n	ai=wa	(**ewulli=ci+anh.ki=l**)]^{NMZ}	[pala=ass=ta]^{FIN}

N=Plur=Top	N=Nom **ACT**	Adj=Part	N=Com	(V+Neg) = Nmz=Acc **Q**	V=Past= Fin.Decl
они	Дадли	такой	ребёнок	не общаться	желать
Они желали, чтобы Дадли не общался с таким ребенком.					

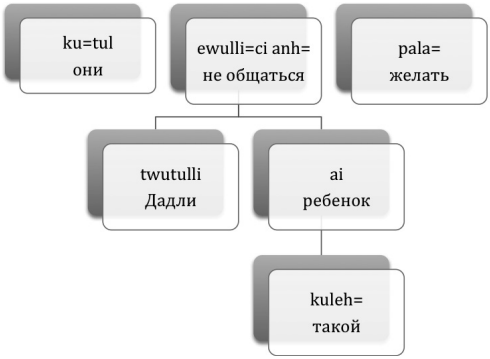


Рис. 3. Act – актант для Q: поверхностно-синтаксическое представление

4. Результаты

Конфигурация вида [Act Part Host] встретилась примерно в 110 предложениях корпуса. Исследованный корпус состоит из текстов двух жанров: художественного (1 глава из перевода книги “Гарри Поттер и философский камень” Джоан Роулинг, общей длиной в 436 предложений) и публицистического (20 коротких новостных и научно-популярных заметок, общей длиной в 200 предложений). В публицистическом жанре, что неудивительно, причастные обороты встречаются почти в два раза чаще.

Ниже представлены условия выбора хозяина для имени Act, сформулированные на основе наблюдений за данными корпуса. Необходимо заметить, что в одном предложении часто выполняются несколько условий одновременно: например, одновременно могут выполняться условия (2) и (3) (см. Табл.8) или условия (2) и (4) (см. Табл.11). Кроме того, каждое конкретное условие действует только на ближайшее к нему слева имя Act, если их несколько.

*4.1. Act – это слуга причастия Par*t

- Условие 1. Топикальная платформа

Или непосредственно Host, синтаксический хозяин причастия Part, несет частицу топика, или между Part и именем с частицей топика нет других предикативных словоформ. В табл.5 и табл.6 представлены иллюстрации для топикальных платформ, в которых имя Act имеет показатели номинатива и аккузатива соответственно и является слугой причастия Part.

Таблица 5

Act=Nom в топикальной платформе

[na=pota=to	**nai=ka**	**manh=un**	ce	**namca=nun**]^{TOP}
N=Compar=Pcl	N=Nom **ACT**	Adj=Part **PART**	N=Ø	N=Top **HOST**
я – даже более чем	возраст	большой	тот	мужчина

[way	eymeylaltupich	choloksayk	mangtho=lul	(ip=ko+iss=nun+kes+i=ci)?]^{FIN}
N=Ø	N=Ø	N=Ø	N=Acc	V=Progress=Fin.Quest
почему	изумрудный	зеленый	мантия	надевать

{Мистер Дурсль пришел в ярость, увидев, что некоторые из них совсем не молоды, – подумать только,} *один из мужчин выглядел даже старше него, а позволил себе облачиться в изумрудно-зеленую мантию!*

Таблица 6

Act=Acc в топикальной платформе

[phothe	i=la=nun]$^{PART\text{-}QUOT}$	[seng=ey	hayli	i=la=nun]$^{PART\text{-}QUOT}$
N=Ø	Copul=Quot. Decl=Part	N=Dat	N=Ø	Copul=Quot.Decl=Part
Поттер	быть	фамилия	Гарри	быть

[**atul=ul**	**twu=n**	**salam=un**]^{TOP}	[(manh=ul+kes+i=ta)]^{FIN}
N=Acc **ACT**	V=Part **PART**	N=Top **HOST**	Adj=Fin.Decl
сын	класть	человек	многий

{Мистер Дурсль легко убедил себя в том, что в Англии} *живет множество семей, носящих фамилию Поттер и имеющих сына по имени Гарри.*

- Условие 2. Host – предикативное имя типа Fact

В позиции Host, синтаксического хозяина причастия Part, располагается одно из класса предикативных имен, например таких, как kes *факт*, kyengwu *случай*, ttay *время*, phil.yo *необходимость*, il *событие*.

Таблица 7

Host – это имя типа Fact.
Одиночное причастие

[i	salam=tul=un]^{TOP}	[**mokum=ul**	(**ha=ko+iss=nun**)	**kes=i**	punmyengha=e…]^{FIN}
N=Ø	N=Plur=Top	N=Acc **ACT**	V=Process.Part **PART**	N=Nom **HOST**	Adj=Fin.Decl
этот	человек	пожертвование	делать	факт	очевидный
{Но тут мистера Дурсля осенила мысль, что} *эти* {непонятные} *личности наверняка всего лишь собирают пожертвования или что-нибудь в этом роде...*					

Интересно, что данные корпуса свидетельствуют, что если в предложении встретилось два или три причастия при одном общем имени-хозяине Host, то левее всех располагается именно то отглагольное причастие, у которого выражены актанты, отличные от Host. Образец этой ситуации представлен в предложении ниже (см. Табл.8).

Таблица 8

Host – это имя типа Fact.
Пара причастий при общем хозяине

[i	ai=nun]^{TOP}	[teculli	pupu=ka	phothe	**pupu=lul**
N=Ø	N=Top	N=Ø	N=Nom	N=Ø	N=Acc **ACT**
этот	ребенок	Дурсль	супруги	Поттер	супруги

melliha=nun	tto	talu=n	**iyu**	(i=ki=to+ha=ess=ta)]^{FIN}
P=Part **PART**	N=Ø	Adj=Part	N=Ø **HOST**	Copul=Fin.Decl
отстранять	еще	другой	причина	быть
букв. Этот ребенок был еще одной причиной, по которой Дурсли отстраняли Поттеров.				

• Условие 3. Финитная и топикальная платформы одновременно

Если конфигурация [Act Part Host] встретилась внутри финитной платформы и при этом в данном предложении есть топикальная платформа, то во всех отмеченных в корпусе случаях имя Act (как с показателем номинатива, так и с показателем аккузатива) оказалось актантом причастия Part.

Таблица 9

Финитная и топикальная платформы

[kuleh=ntey]^{GER}	**[ku=nun]^{TOP}**	[ppangcip=eyse	khetalah=n	tones
Adj=Ger	N=Top	N=Abl	Adj=Part	N=Ø
но	он	булочная	большой	Донатс
pongci=lul	(tul=ko+nao=taka)]^{GER}	[wuyenhi	**ku=tul=i**	**cwukopat=nun**
N=Acc	V=Ger	Adj=Adv	N=Plur=Nom **ACT**	V=Part **PART**
пончик	держать+выходить	случайно	они	обмениваться
myech	mati=lul	(tut=key+toy=ess=ta)]^{FIN}		
N=Ø	N=Acc	V=Past=Fin.Decl		
несколько	слова	слышать		
Выйдя из булочной с пакетом, в котором лежал большой пончик, […] *и в этот момент он абсолютно случайно услышал* […].				

• Условие 4. Совпадение актантных показателей на именах Act и Host

Это условие может не выполняться в случае глаголов, допускающих биноминативную конструкцию, например, глагола toy= *становиться.*

Таблица 10

Два аккузатива: Act=Acc Part Host=Acc

[teculli	ssi=nun]^{TOP}	[celm.un.ay=tul=ina	ip=nun	(isangha=ko+wusukkwangsule=n)
N=Ø	N=Top	N=Plur=Pcl	V=Part	Adj=Part
Дурсль	господин	молодой человек	носить	странный+смехотворный

os=ul	**ip=un**	**salam=tul=ul**	po=myen]^{GER}	[(cham=ul+swu=ka+ eps=ess=ta)]^{FIN}
N=Acc **ACT**	V=Part **PART**	N=Plur=Acc **HOST**	V=Ger	V=Past=Fin.Decl
одежда	носить	люди	смотреть	терпеть + не мочь
Мистер Дурсль не переносил людей в нелепой одежде, да взять хотя бы нынешнюю молодежь, которая расхаживает чёрт знает в чем!				

Таблица 11

Два номинатива: Act=Nom Part Host=Nom

[onul=un]^{TOP}	[achimhay=ka	tteolu=n+ihwu =lo]^{GER}	[swupayk	meli=uy	**puengi=ka**
N=Top	N=Nom	V=Part+N=Ins	N=Ø	N=Gen	N=Nom **ACT**
сегодня	утреннее солнце	подниматься	несколько сотен	голова	сова

sapang=eyse	**nal.atani=nun**	**kwangkyeng=i**	pel.eci=ess= supnita]^{FIN}
N=Abl	V=Part **PART**	N=Nom **HOST**	V=Past=Fin.Decl
4 стороны света	летать туда-сюда	зрелище	открываться
[…] *сегодня поступали сотни сообщений от людей, которые с самого рассвета в разных точках страны видели беспорядочно летающих сов.*			

4.2. Act – это слуга Q, главного предикатива платформы

Все отмеченные в корпусе примеры этого класса имеют отношение только к именам с показателем номинатива. При этом рассматриваемое имя не должно подпадать ни под одно из условий выше. Примеры этого сорта встречаются среди предложений корпуса довольно редко.

- Условие 5. Нефинитная платформа

Таблица 12

Act и Part в нефинитной платформе

[**ku=ka**	cheum=ulo	mwe=inka	com	**isangha=n**	kkimsay=lul
N=Nom **ACT**	N=Ins	N=Pcl	N=Ø	Adj=Part **PART**	N=Acc
он	впервые	что-то	немного	странный	симптом

[**al.achali=n**	ke=n]$^{\mathbf{TOP}}$	[tolo	mothwungi=lul	mak	tol.ase=ess=ul+ttay]$^{\mathbf{GER}}$	i=ess=ta]$^{\mathbf{FIN}}$
V=Part **Q**	N=Top	N=Ø	N=Acc	N=Ø	V=Past=Part+N	Copul=Fin. Decl
осознавать	факт	обратно	поворот	резко	повернуть и остановиться+время	быть
букв. Он впервые заметил, что происходит что-то странное, когда резко повернул за угол и затормозил.						

Аналогичная ситуация имеет место во второй платформе в (Табл.4) выше.

Интересно, что нефинитная платформа часто соседствует с топикальной, так что в непосредственном соседстве оказывается словоформа с показателем Top и рассматриваемая словоформа с показателем Nom из нефинитной платформы. См. пример **(8)** выше.

- Условие 6. Финитная платформа при отсутствии топика

Таблица 13

Act и Part в финитной платформе

[hiasinsu	**hyangki=ka**	**tallyeka=nun**	kunye=lul	**ttalao=ass=ta**]$^{\mathbf{FIN}}$
N=Ø	N=Nom **ACT**	V=Part **PART**	N=Acc	V=Past=Decl.Fin **Q**
гиацинт	аромат	убегать	она	догонять
Ее, убегавшую, преследовал аромат гиацинтов.				

REFERENCES

Martin, S. E. (1992). *A reference grammar of Korean*. Rutland, Vermont; Tokyo: C.E.Tuttle.

Brechalova, E. (2010). On the procedure of formal analysis of Korean text. *Cahiers d'études coréennes, Num.8* (pp. 27–44). Paris: Institute d'études coréennes, College de France.

Бречалова, Е.В. (2008). *Принципы построения синтаксического представления корейского предложения*. Дисс. на соискание степени к.фил.н.. Москва.

THE LOCATIVE ATTRIBUTIVE OF THE TATAR LANGUAGE: THE STRUCTURE AND SEMANTICS

Alfiya Galieva

Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences,
Kazan, Tatarstan, Russia
amgalieva@gmail.com

Due to significant structural specificity of the Turkic languages, having agglutinative morphology, there is no strict distribution of words and word forms within grammatical classes and parts of speech. The paper deals with Tatar word forms, containing affix of the locative attributive DAGI; and the structural and semantic peculiarities of these forms of hybrid nature are examined.

Using data from "Tugan Tel" Tatar National Corpus, the author reveals and describes typical affixal environment of the DAGI affix in real texts (the types of stems are distinguished, the affixes to the left and to the right of the DAGI affix are described, and typical affixal chains are considered). Grammatical case affixes, and forms of Comparative, Similative and Equative are of special interest. Statistical information on distribution of different types of word forms having DAGI affix is given. For the proper interpretation of the word form, containing affix DAGI, the elements of the affixal chain are important as well as order of their disposition. The use of the locative attributive enables to introduce phrases and dependent clauses of attributive meaning into the sentence.

The analysis justifies the idea that the affix of the locative attributive DAGI is an affix, intermediate between inflectional and word formative affixes. The affix may be used in in rich variety of affixal chains and it favours denoting complicated relations between the objects of the world.

1. Введение

В силу значительного структурного отличия тюркских языков от индоевропейских, обусловленных агглютинацией, в тюркских языках наблюдается размывание границ парадигматических классов при потенциально неограниченном объеме парадигмы и нежесткое распределение лексики по грамматическим классам и частям речи. Граница между словообразованием и словоизменением во многих случаях также проводится весьма условно, так как многие аффиксы интерпретируются как словоизменительные или деривационные в зависимости от структуры аффиксальной цепочки или контекстуального окружения того или иного обра-

зования. Одним из самых примечательных аффиксов такого рода, промежуточных на границе словообразования и словоизменения, является аффикс -ГЫ (Щербак, 1977), который в современном татарском языке может функционировать как самостоятельно, так и в составе сложных аффиксов, в том числе аффикса ДАГЫ.

В статье анализируются структурно-семантические особенности гибридных образований, формируемых путем присоединения аффикса локативного аттрибутива ДАГЫ к основам разного типа. Описывается состав и структура словарной цепочки слева и справа от анализируемого аффикса. Языковой материал извлечен из Татарского национального корпуса «Туган тел» (corpus.antat.ru).

Аффикс ДАГЫ ряд исследователей (М.З. Закиев, Ф.А. Ганиев, Р.Ф. Зарипов) включает в состав падежных аффиксов, основываясь на том, что этот аффикс устанавливает связь существительного с другими словами (Татарская грамматика, 1993, т. 2). Для подтверждения или опровержения этого тезиса требуется тщательное исследование структуры и семантики образований на ДАГЫ с применением корпусной коллекции текстов, настоящее исследование является первым шагом в этом направлении.

2. Основные способы формирования образований на ДАГЫ

Вначале рассмотрим основные способы выражения определительных отношений в татарском языке.

1. Порядок слов, когда определяющее слова располагается слева от определяемого.

В тюркских языках отсутствует какое-либо согласование между компонентами атрибутивных конструкций, и отсутствуют словоизменительные категории, которые могли бы инициировать такое согласование (Гузев, Бурыкин, 2007). Порядок слов характеризует определительные конструкции типа *прилагательное* + *существительное* и *существительное* + *существительное* (изафет 1).

Изафет II отличается наличием аффикса принадлежности у определяемого компонента и служит средством выражения относительного признака; изафет III представляет собой конструкцию, в которой компонент-определение выступает в форме родительного падежа, а определяемое – в форме принадлежности 3-го лица,

и служит средством передачи предметных отношений, воспринимаемых говорящим как притяжательные (Гузев, Бурыкин, 2007).

Кроме этого, имеются специальные аффиксы, преобразующие существительные в прилагательные и позволяющие образовать атрибутивные конструкции разного типа.

Аффикс ДАГЫ представляет собой сложный аффикс, образованный присоединением релятивизатора ГЫ к аффиксу локатива.

Аффикс ГЫ преобразует существительные, наречия и глаголы в прилагательные со значением относительного признака (Татарская грамматика, 1993. т.1):

көз – көзге көн (=изафет *көз көне*);

көндез – көндезге аш (изафет **көндез аш(ы)* аграматичен).

Таким образом, аффикс ГЫ позволяет создавать атрибутивы и в тех случаях, когда имеются ограничения на образование атрибутивных словосочетания на базе изафета.

Локативный аттрибутив передает главным образом пространственные (1, 2) и, несколько реже, временные (3) отношения:

(1) *авылдагы бабай*
деревня-ATTR_LOC дед
дед, который в деревне;
(2) *Казандагы кызы*
Казань- ATTR_LOC дочь-POSS_3
дочь в Казани;
(3) *алтыдагы автобус*
шесть- ATTR_LOC автобус
автобус, который в шесть.

Аффикс ДАГЫ присоединяется к основам разного типа:

- к существительным (4–6):

(4) *китаптагы*
книга- ATTR_LOC
который в книге;
(5) *Казандагы*
Казань- ATTR_LOC
который в Казани);
(6) *зурлыктагы*
величина- ATTR_LOC
величиной с);

- местоимениям и местоименным наречиям (7–10):

(7) *миндәге*
я-ATTR_LOC
который у меня;
(8) *алардагы*
они- ATTR_LOC
который у них;
(9) *кайдагы*
где-ATTR_LOC
который где;
(10) *биредәге*
здесь- ATTR_LOC
который здесь;

- прилагательным и наречиям (11–12):

(11) *якындагы*
близкий- ATTR_LOC
который поблизости;
(12) *түбәндәге*
низкий- ATTR_LOC
который внизу;

- числительным (13):

(13) *биштәг*
пять- ATTR_LOC;
который в пять;
именам действий (14–15):
(14) *артудагы*
увеличение – ATTR_LOC
который в росте;
(15) *үстерүдәг*
выращивание- ATTR_LOC
который при выращивании;

- причастиям (16–17)*:*

(16) *кайткандагы*
вернуться-PTC_PS, -ATTR_LOC
который при возвращении;
(17) *алгандагы*
брать-PTC_PS, ATTR_LOC
который при взятии.

Образования на ДАГЫ могут функционировать как послеложные слова, замещая послелоги в составе определительных конструкций: *турындагы* (о), *хакындагы* (о), *арасындагы* (между), *буендагы* (вдоль) и т.п.:

сугыш турында (о войне) – сугыш турындагы китап (книга о войне);

бәйләнеш хакында (о связи) – бәйләнеш хакындагы фикер (мысль о связи);

егет белән кыз арасында (между девушкой и юношей) – *егет белән кыз арасындагы мөнәсәбәт* (отношения между девушкой и юношей);

юл буенда (вдоль дороги) – *юл буендагы шомырт* (черемуха вдоль дороги).

3. Структура аффиксальной цепочки образований на ДАГЫ

3.1. Структура аффиксальной цепочки слева от аффикса локативного атрибутива

Слева от аффикса локативного атрибутива могут стоять словоизменительные аффиксы разных типов:

1. Аффикс множественного числа ЛАр:

(18) *рус-лар-дагы*

русский-PL, ATTR_LOC

тот, что у русских;

(19) *ел-лар-дагы*

год-PL, ATTR_LOC

тот, что в годы

2. Аффиксы принадлежности (наиболее часто встречается аффикс принадлежности 3-ему лицу):

(20) *состав-ы-ндагы*

состав-POSS_3, ATTR_LOC

тот, что в составе (чего-л.);

(21) *зурлыг-ы-ндагы*

величина-OSS_3, ATTR_LOC

величиной (с чем-л.).

3. Аффикс компаратива:

(22) *ераг-рак-тагы*
далеко-COMP, ATTR_LOC
тот, который подальше;
(23) *югары-рак-тагы*
высоко-COMP, ATTR_LOC
тот, который повыше
Часто встречаются комбинации аффиксов разного типа (24–25):
(24) *як-лар-ы-ндагы*
Сторона- PL, POSS_3, ATTR_LOC
тот (те), что по сторонам;
(25) *теге-ндә-рәк-тәге*
тот-LOC, COMP, ATTR_LOC
тот, который там подальше
В случае, если аффикс ДАГЫ является крайним справа, образование на ДАГЫ функционирует как атрибутив, поясняющий значение субстантива справа:
(26) *теге-ндә-рәк-тәге каен*
тот-LOC, COMP, ATTR_LOC береза
береза, та, которая там подальше.

3.2. Структура аффиксальной цепочки справа от аффикса локативного атрибутива

При субстантивации или адвербиализации образования на ДАГЫ справа от аффикса локативного атрибутива также могут стоять аффиксы разных типов:
1. Аффикс множественного числа ЛАр:
(27) *өй-дәге-ләр*
дом-LOC_ATTR, PL
те, которые дома;
(28) *кием-дәге-ләр*
одежда-LOC_ATTR, PL
те, которые в одежде.
Расположение аффикса множественного числа по отношению к аффиксу ДАГЫ меняет значение словоформы (29–30)
(29) *өйләрдәге*
дом-PL, ATTR_LOC

тот, который в домах;
(30) *өйдәгеләр*
дом-ATTR_LOC, PL
те, которые в доме
2. Аффиксы косвенных падежей.

Согласно корпусной коллекции, почти все случаи употребления косвенных падежей в составе образования на ДАГЫ – с аффиксом множественного числа:

(31) *заводтагыларның*
завод-ATTR_LOC, PL, GEN
(у) тех, которые на заводе;
(32) *йорттагыларга*
дом-ATTR_LOC, PL, DIR
тем, которые дома;
(33) *тоткындагыларны*
плен-ATTR_LOC, PL, ACC
тех, которые в плену.

Таблица 1 представляет распределение локативных аттрибутивов, осложненных аффиксами падежей, в коллекции Татарского национального корпуса «Туган тел». Аффикс падежа может стоять как после, так и перед аффиксом ДАГЫ.

Таблица 1

Количество форм на ДАГЫ, имеющих падежные аффиксы

Косвенные падежи	Количество форм на ДАГЫ
генетив	170
датив	277
аккузатив	553
аблатив	79
локатив	39

3. Аффикс экватива -чА:
(34) *тормыштагыча*
жизнь-ATTR_LOC, EQU
как в жизни;

(35) *түбәндәгечә*
нижний-ATTR_LOC, EQU
как (написано) ниже;
(36) *бездәгечә*
мы-ATTR_LOC, EQU
по-нашему (как у нас).

В ряде случаев образование на ДАГЫ с аффиксом экватива может быть осложнено дополнительно аффиксом компаратива, это происходит обычно в случаях, когда локативный атрибутив определяет глагол (37–39) и, реже, существительное:

(37) *җирдәгечәрәк булып чыкты*
земля-ATTR_LOC, EQU, COMP быть-ADVV выходить-PST_DEF
оказалось немного так, как бывает на земле;
(38) *җавап түбәндәгечәрәк булды*
ответ нижний-ATTR_LOC, EQU, COMP быть-PST_DEF
ответ был похож на тот, что ниже;
(39) *түбәндәгечәрәк яңгырый*
низкий/нижний-ATTR_LOC, EQU, COM звучать-PRES
звучит немного так, как (указано) ниже

4. Аффикс симилятива ДАй:
(40) *сәхнәдәгедәй*
сцена-ATTR_LOC, SIM_1
словно на сцене
(41) *бездәгедәй*
мы-ATTR_LOC, SIM_1
словно у нас

В таблице 2 представлено количество употреблений образований с аффиксом ДАГЫ, осложненных аффиксом множественного числа, экватива и симилятива.

Таблица 2

Количество форм на ДАГЫ, имеющих аффиксы множественного числа, экватива и симилятива

Всего	С аффиксом множественного числа	Саффиксом экватива	С аффиксом экватива и компаратива	С аффиксом симилятива
36 000	4 000	650	4	73

Образование на ДАГЫ часто вводит не отдельное слово, а целое словосочетание (42–43):

(42) *олы яшьләрдәге бер агай*

большой год-PL, ATTR_LOC один дядя

один пожилой мужчина

(43) *Идел буе районындагы*

Волга длина-POSS_3, район-POSS_3, ATTR_LOC

тот, который в Приволжском районе.

Образования с симилятивом Дай в большинстве случаев вводят не отдельные слова, а словосочетания (44) или придаточные предложения (45):

(44) *уч төбендәгедәй җәелеп*

ладонь дно-ATTR_LOC, SIM_1 простираться-ADVV

простираясь, как на ладони;

(45) *Камилнең колаклары корт күче чаккандагыдай шаулый.*

Камиль-GEN ухо-PL, POSS_3 пчела рой--POSS_3 жалить-PCP_PS,ATTR_LOC, SIM_1 шуметь-PRES

В ушах у Камиля звенело, словно его ужалил целый пчелиный рой.

Заключение

Анализ корпусных данных позволяет придти к заключению, что аффикс локативного атрибутива активно используется в татарском языке в аффиксальных цепочках, отличающихся значительным разнообразием. Аффикс ДАГЫ может стоять комбинациях с большим количеством других аффиксов в пре- и постпозиции к аффиксу ДАГЫ (аффиксы падежей, множественного числа, компаратива, экватива, локатива и др.). Для интерпретации словоформы на ДАГЫ большое значение имеет как сами аффиксы, так и порядок их расположения. Локативный атрибутив позволяет вводить как отдельные словосочетания, так и придаточные предложения, выражающие очень сложные связи между объектами мира.

Проведенная работа подтверждает тезис о том, что в языках аглютинативного типа аффиксы не имеют четкого деления на словообразовательные и словоизменительные, но цепочки аффиксов имеют строгие правила последовательного присоединения элементов в зависимости от типа словоформы.

ЛИТЕРАТУРА

Щербак А.М. (1977). *Очерки по сравнительной морфологии тюркских языков (имя)*. – М.: Наука.

Гузев В.Г., Бурыкин А.А. (2007). *Общие строевые особенности агглютинативных языков* //Acta linguistica Petropolitana. Труды ИЛИ РАН. – Т. 3. Ч. 1 (с. 109–117) СПб.

Татарская грамматика: (1993). В 3 т. Т. 1. Введение. Фонетика. Фонология / Рос.АН, АН Татарстана, Ин-т яз., лит. и ист. им. Г. Ибрагимова; Релкол.: М.З. Закиев (гл. ред.) и др. – Казань: Татар. кн. изд-во.

Татарская грамматика: (1993). В 3 т. Т. 2. Морфология / Рос. АН, АН Татарстана, Ин-т яз., лит. и ист. им. Г. Ибрагимова; Казан. Науч. центр; Редкол.: М.З. Закиев и др.. – Казань: Татар. кн. Изд-во.

Татарский национальный корпус «Туган тел» // URL: corpus.antat.ru.

STATISTIC DISTRIBUTION OF SOME GRAMMATICAL CATEGORIES OF THE TATAR LANGUAGE OVER CORPUS DATA

Alfiya Galieva [a], Olga Nevzorova [a, b], Dzhavdet Suleymanov [a, b]

[a] Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Levo-Bulachnaya, 36a, Kazan, 420111, Russia,
[b] Kazan Federal University, Kremlevskaya, 18, Kazan, 420000, Russia
[1]amgalieva@gmail.com

The main objective of the paper is to discuss the construct of linguistic complexity in Tatar and to show how it may be measured in Tatar language processing applications. The Tatar language is an agglutinative language with a highly productive complicated inflectional and derivational morphology. We consider the statistical distribution of grammatical categories on corpus data and emphasize the complex phenomenon of Tatar morphology. The paper shows that Tatar inflectional affixes tend to be universal for different parts of speech, and all the allomorphs are conditioned phonetically, and irregular affixes do not exist. We demonstrate also the frequency of allomorphs and give some explanations concerning asymmetry of distribution of allomorphs. Particular attention is paid to nominal affixes, we show statistically that Tatar case affixes and the comparative affix are not bound to the unique grammatical class of stem, but may be joined to a broad range of stems.

Linguistic data reflects the current state of Tatar National corpus.

1. Introduction

The main objective of this paper is to discuss the construct of linguistic complexity in the Tatar language and to show how it is typically measured in Tatar language processing applications.Currently linguistic complexity is one of debatable notions in linguistics, and there are different ways of understanding complexity, depending on linguistic domains and researcher's aims:

- linguistic complexity in second language acquisition;
- linguistic complexity in contrastive linguistics and typology;
- linguistic complexity in natural language processing, etc.

Evidently, quantitative properties of the language are essential for describing linguistic complexity.

From our viewpoint, among the factors that influence the complexity of a language may be viewed the following:

- universality of means for expressing grammatical categories (or its absence);
- variety of grammatical categories of different types;
- regularity of means of expression of a grammatical category (absence of exception to the rules);
- degree of linguistic redundancy;
- length of affixal chains (average length of affixal chains for each part of speech);
- potential of interconversion of parts of speech, etc.

The task of calculating the frequency of grammatical categories and distribution of allomorphs in real texts is the first step in the assessment of linguistic complexity.

The paper represents an endeavour to assess the morphological complexity of the Tatar language as a first approximation, by analysing the distribution of certain grammatical categories on corpus data.

In the paper we sketch out some features for the assessment of the linguistic complexity of the Tatar language relevant for natural language processing.

The paper is organized as follows: in Section 2 we sketch the background of the research; in Section 3 we outline some typological features of the Tatar language; Section 4 presents the distribution of Tatar affixes on the data of part-of-speech-tagged Tatar National Corpus.

2. Related work

Over the last three decades quantitative methods are constantly gaining importance in all branches of linguistics and NLP. The quantitative approach to the language opens up new theoretical perspectives and proposes solutions to a wide range of practical problems.

The International Quantitative Linguistics Association (IQLA) was founded in 1994 (International Quantitative Linguistics Association. URL: http://www.iqla.org). Corpus technologies enable researchers to obtain exact information on the distribution of linguistic units and grammatical categories. Languages of rich morphology are of special interest for us.

The paper by M.V. Kopotev et al. (Kopotev, 2008) presents an approach to building frequency grammar of Russian and describes the distribution of grammatical cases regarding parts of speech and genres of texts.

Frequency and declensional morphology of Czech nouns is presented in Mačutek & Čech (Mačutek & Čech, 2012), where the relationship among frequency, morphology and phonetics is observed.

Special literature outlines the most important phonological, morphological and syntactic features of the Turkic languages. K. Oflazer presents the main characteristics of Turkish and some challenges for language processing (Oflazer, 2014). Many papers are devoted to morphological disambiguation of Turkic languages, especially Turkish (Daybelge & Çiçekli, 2007, Hakkani-Tür, Oflazer, &Tür, 2002, Sak, Güngör & Saraçlar, 2008, Oflazer & Kuruöz, 1994).

Tatar grammars (Tatar Grammar, 1993, Tatar Grammar, 2002) give basic description of the grammatical structure of the Tatar language, however they do not contain any information on statistical distribution of the described phenomena. Developed corpora of Tatar may provide data for reliable language analysis, enabling an accurate description of lexis and grammar. Nevertheless serious quantitative analysis of corpus data has not been implemented yet.

Brief statistical information on the system of the Tatar language is presented on the website of the Corpus of Written Tatar (Corpus of Written Tatar. URL: http://corpus.tatfolk.ru/index_en.php.), developed at Kazan Federal University, with a list of the most frequent word forms, a list of most frequent n-grams (word forms and combinations of letters in different positions in a word, etc.). For our statistical experiments we have used the new Tatar National Corpus (Tatar National Corpus. URL: http://corpus.antat.ru, Suleymanov, Nevzorova, Gatiatullin, Gilmullin & Khakimov, 2013).

3. Typological features of the Tatar language

The Tatar language belongs to the Turkic group that forms a subfamily of the Altaic languages. The Tatar language is spoken in west-central Russia (in Volga region) and southern parts of Siberia. The number of Tatar speakers in Russia in 2010 was 4.28 million people.

The most important phonetic feature of the Turkic languages is progressive vowel harmony.

The basic way of word formation and inflection is progressive affixal agglutination when a new unit is built by consecutive addition of regular and clear-cut monosyllabic derivational and inflectional affixes to the stem, therefore the stem remains unchanged. Affixal agglutination

provides unified morphological means for forming derivatives within the same grammatical class of words as well as for changing the part-of-speech characteristic of the word and for turning it into another lexical and grammatical class. The boundaries between the affixes within the word form are distinct and transparent, and the affixal joint in many cases coincides with the syllabication (Guzev, Burykin, 2007).

The order of affixes is rigidly determined, and derivational affixes (suffixes) precede inflectional ones. Each added suffix tends to modify the whole preceding stem.

Words have no classifying categories (like grammatical gender and animacy). The affixes in the affixal chain are mainly unambiguous. There is one type of declension and conjugation, and one set of affixes is used only. The Tatar language has no grammatical prefixes and prepositions, although it has postpositions.

In Tatar, as in a language of agglutinative structure, the paradigm as a set of ready-made word forms is absent (because of the limitlessness of paradigms), and word forms are generated directly in the act of speaking:

(1) *kitap-lar-ı-nan-mı*
book-PL, POSS_3SG, ABL, INT
'whether from his books', 'from his books?'

Another typologically relevant feature of the Tatar language is absence of clear-cut borders between inflection and word formation, since the same affixes in different positions may function both as inflectional and as derivational.

Tatar nouns are marked with regard to their number, possession and case and are not characterised by definiteness. The Tatar verb has no aspect, but is characterised by tense, mood and can have the negative form.

The plural of nouns is formed by joining the affix –LAr to the stem; the same plural affix is used to substantivize adjectives and to form 3d person of verbs. Possessive affixes are used to express the person and the number of possessors.

An example of the word form below represents some salient features of Tatar morphology:

(2) *Qadaq-la-š-qan-nar-ı-na*
Qadaq -a nail, a tack (NOUN)
Qadaq-la – to nail (VERB)
Qadaq-la-š – to help to nail (VERB)

Qadaq-la-š-qan –[he, she] helped to nail (VERB, PARTICIPLE)
Qadaq-la-š-qan-nar- [they] helped to nail (VERB, PARTICIPLE)
Qadaq-la-š-qan-nar-ı – those who helped to nail (NOUN)
Qadaq-la-š-qan-nar-ı-n – to those who helped to nail (NOUN)

Tatar Morphology is regular and predictable in many respects, and there is little or no fusion between the stem and the affixes.

4. Tatar agglutinative morphology in corpus data

4.1. Distribution of nominal inflectional affixes

In this section, we will present statistical results about the corpus in order to get an idea about the coverage of a corpus of this size for an agglutinative language and also to observe the morphological characteristics of Tatar language.

The statistical distribution of grammatical categories were derived from Tatar National Corpus (Tatar National Corpus. URL: http://corpus.antat.ru). The Tatar Corpus includes mainly prosaic texts which represent the literary Tatar language (from the 20th of the XIX century in Cyrillic alphabet), as well as modern scientific and business texts, texts of official documents and newspaper materials. All the texts included in the Tatar corpus go through special procedures of meta-annotation (attributing of metadata to the text) and morphological annotation (attributing of morphological information to each word-form). Now Tatar National Corpus includes about 46 M word forms.

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. For example, there is only one special means for expressing plurality, both for nouns and verbs:

(3) *Bala-lar bara-lar.*
child-PL go- PRES, PL
children go

The same affix may be used in cases when adjectives related to nouns in plural, serve as predicates:

(4) *Alma-lar qızıl-lar.*
apple-PL red-PL
Apples are red.

Figure 1 shows the distribution of affix of plurality (PL) on data from Tatar National corpus).

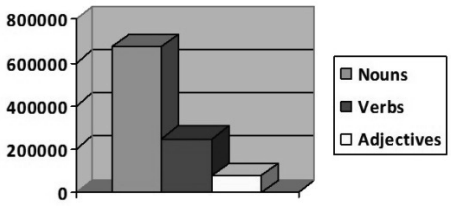


Fig. 1. Corpus distribution of the morpheme of plurality (PL)

The language tend to avoid redundancy, and so in many cases, if plural affix is used in noun, it is missed out or may be missed out in related words – verbs and adjectives.

Although the plurality meaning component is present in boundry in forms of 1st and 2nd persons plural of pronouns and verbs, Tatar morphology does not mark it as standard plural (we is not plural of I, and You (Pl) is not plural of you (Sg)).

Tatar has only one type of declension for all parts of speech, as well as one type of verb conjugation (there are no verbs of irregular conjugation).

The number of grammatical cases in Turkic languages is a subject of ongoing discussion (Tatar Grammar, 1993, pp. 42–45). Tatar grammars traditionally distinguish 6 cases: Nominative (unmarked case), Genitive, Directive, Accusative, Ablative and Locative [9, 10]. The developers of the "Tugan Tel" Tatar National Corpus distinguish one more case – Directive with a limitative meaning (Tatar National Corpus. URL: http://corpus.antat.ru). Table 1 represents the frequency of case affixes depending on stem type.

Table 1

Distribution of case affixes in parts of speech

Case	Total	Word Stem			
		Noun	Adjective	Verb	Other parts of speech (in sum)
NOM	3597000	2352000 (N,NOM)	1245000 (ADJ,NOM)		
GEN	483000 (GEN)	363000 (N,GEN)	14000 (ADJ,GEN)	31000 (V,GEN)	202000 (!(N|V|ADJ),GEN)

DIR	948000 (DIR)	671000 (N,DIR)	120000 (ADJ,DIR)	346000 (V,DIR)	250000 (!(N|V|ADJ),DIR)
ACC	1201000 (ACC)	959000 (N,ACC)	158000 (ADJ,ACC)	205000 (V,ACC)	540000 (!(N|V|ADJ),ACC)
ABL	314000 (ABL)	251000 (N,ABL)	62000 (ADJ,ABL)	60000 (V,ABL)	29000 (!(N|V|ADJ),ABL)
LOC	604000 (LOC)	507000 (N,LOC)	71000 (ADJ,LOC)	174000 (V,LOC)	50000 (!(N|V|ADJ),LOC)

Affixes of indirect cases join word stems of different classes: nouns, substantivized adjectives, verbal stems (deverbal nouns, participles, deverbal adjectives of different types), pronouns, etc. Although the number of nominal stems prevails over other types, the amount of verbal derivatives, joining case affixes, is also significant. Examples (5-9) represent use of Locative affix joining stems of different types: nominal (5, 6), adjectival (7), verbal (8, 9).

(5) *urman-da*
forest-LOC
'in forest'

(6) *al-da*
front-LOC
'in front of, in advance of'

(7) *qara-da*
black-LOC
'in black'

(8) *qayt-ma-w-da*
return-NEG, DV, LOC
'in not returning'

(9) *qayt-qan-da*
return-PCP_PS, LOC
'when (sb) returns'

Example (10) illustrate a real use of (9):

(10) *Qaytkan-da adašmassıŋmı soŋ?*
return-PCP_PS, LOC lose one's way-FUT_IND_NEG, 2SG, INTR indeed
Won't you lose your way back?

A relatively small number of cases when the Genitive is used with verbal derivatives may be explained by the fact that the Turkic genitive forms possessive attributes and has certain constraints for combining with verbal derivatives.

4.2. Distribution of allomorphs (case affixes)

As a result of progressive vowel harmony the quality of second and subsequent syllables is determined by the quality of the first or preced-

ing syllable, and certain researchers claim that second and subsequent syllables of Turkic languages may contain three phonemes (or morphophonemes) only - a, ı, u (Guzev, Burykin, 2007).

Table 2

The phonetic aspect of the stem

Cases	Affixes of back vowel	Example of word form / English translation	Affixes of front vowel	Example of word form / English translation
nominative	ø	qul / 'hand'	ø	kül / 'lake'
genetive	-nıŋ	qulnıŋ /'of hand'	-neŋ	külneŋ / 'of lake'
directive	-ga	qulga / 'to hand'	-gä	külgä / 'to lake'
accusative	-nı	qulnı / 'hand'	-ne	külne / 'lake'
ablative	-dan	quldan / 'from hand'	-dän	küldän / 'from lake'
locative	-da	qulda / 'in hand'	-dä	küldä / 'in lake'

The Tatar language exhibits complete stem invariance; depending on the phonetic aspect of the stem, affixes of two vowel types are used (Table 2). Quality consonants in case affixes are also determined by the character of the last sound of the stem: in the directive, ablative and locative (Table 3). Table 4 represents the distribution of phonological variations for case affixes.

Table 3

The quality consonants in case affixes

Cases	The last sound of the stem is a vowel or a voiced consonant	Qar 'snow'	The last sound of the stem is a voiceless consonant	Taŝ 'stone'	The last sound of the stem is a nasal consonant	Qan 'blood'
directive	-ga (-na)	qarga	-qa	taŝqa	-ga	qanga
ablative	-dan	qardan	-tan	taŝtan	-nan	qannan
locative	-da (-nda)	qarda	-ta	taŝta	-da	qanda

Table 4

Distribution of allomorphs of case affixes

Cases	The stem has a back vowel	The stem has a front vowel	The affix begins with a voiced consonant	The affix begins with a voiceless consonant	The affix begins with a nasal consonant
GEN	-nıŋ 253120	-neŋ 286348	–	–	-nıŋ, -neŋ 539468
DIR	-ga, -qa, -na, -a 614674	-gä, -kä, -nä, -ä 559975	-ga, -gä 635893	-qa, -kä 213627	-na, -nä 325129
ACC	-nı 291562	-ne 263749	–	–	-nı, -ne 555311
ABL	-dan, -tan 124887	-dän, -tän 109008	-dan, -dän 149248	-tan, -tän 84647	-nan, -nän 140449
LOC	-da, -ta 443467	-dä, -tä 333913	-da, -dä 622490	-ta, -tä 154890	-nda, -ndä 397000

On the whole, a number of back and front vowel allomorphs are nearly equal.

The asymmetry between the amount of voiced consonant affix variations and voiceless consonant ones may be explained by the fact that:

- the use of possessive affixes preceding case affixes influence the choice of voiced consonant in case affixes;
- most verbal derivatives in the active voice also phonetically require a voiced consonant in case affixes.

As a result, in the directive, the number of voiced consonants is 2.9 times more than the number of voiceless consonants; in the locative, the number of voiced consonants is 4 times more than that of voiceless consonants.

Nasal consonants in directive, ablative and locative cases are also morphonologically conditioned alternations, they occur mainly in combination of possessive affixes (3SG and 3Pl), which are used to express person and number of possessors. Table 5 represents when nasal variants of case affixes are used.

Table 5

Nasal consonants in directive, ablative and locative cases

Case	Stem+case	Stem+possessive+case	Stem+plural+possessive+case
NOM	Urman ø / ‘forest, wood’	Urman-ı ø / ‘his/her forest'	Urman-nar-ı ø / ‘their forest(s)’
DIR	Urman-ga / ‘to forest’	Urman-ı-na / ‘to his/her forest’	Urman-nar-ı-na / ‘to their forest(s)’
ABL	Urman-nan / ‘from forest’	Uaman-ı-nnan / ‘from his/her forest’	Urman-nar-ı-nan / ‘from their forest(s)’
LOC	Urman-da / ‘in forest’	Urman-ı-nda / ‘in his/her forest’	Urman-nar-ı-nda / ‘in their forest(s)’

4.3. Distribution of Comparatives

Comparatives in Tatar are formed by means of affix -rAQ.

The marker of comparison in Tatar joins not only adjectives, but also case forms of nouns, converbs, participles, adverbs and pronouns. Examples (11, 12) demonstrate the use of comparative affix in verbal word forms:

(11) *qara-b-raq*
look-CONV, COMP
'looking more (attentive)'

(12) *äyt-micä-räk*
say-CONV_NEG, COMP
'more silently'

Comparative affix joins only case forms expressing spatial relations – Directive, Ablative, Locative (and does not join Genetive and Accusative), see examples (13-15):

(13) *qoyaš-qa-raq*
sun-DIR, COMP
'to a sunnier (place)'

(14) *qoyaš-ta-raq*
sun-LOC, COMP
in/at a sunnier (place)

(15) *qoyaš-tan-raq*
sun-ABL, COMP
'from sunnier (place)'

Table 6 represents frequency of comparative affix depending on stem type.

Table 6

Distribution of comparative affix

Word structure	ADJ+ COMP	ADJ + case + COMP			Noun + case + COMP		
		ADJ+ DIR+ COMP	ADJ+ ABL+ COMP	ADJ+ LOC+ COMP	Noun+ DIR+ COMP	Noun+ ABL+ COMP	Noun+ LOC+ COMP

Number of use	55 000	4500	889	2000	1570	690	2660

So Tatar inflectional affixes gravitate towards universality, being the same for different pars of speech. All the allomorphs are conditioned phonetically, and irregular affixes do not exist. The order of affixes is rigidly determined, and that makes meaning of a word form lucid and predictable.

5. Conclusion and future work

The sets of grammatical categories in languages may vary significantly, and quantitative analysis enables researchers to reveal the essence of the language, by revealing the structures and properties we can observe in real texts. We made a first attempt to elucidate certain typological features of the Tatar language from the viewpoint of linguistic complexity.

The paper presented the distribution of nominal inflectional affixes in Tatar on the data of the Tatar National Corpus. We showed that Tatar case affixes and the comparative affix are not bound to the unique grammatical class of stem, but may be joined to a broad range of stems. We also demonstrated the frequency of allomorphes and gave some explanations concerning the asymmetry of distribution of these allomorphes.

The work may be continued in the following directions:

- study of distribution of nominal inflectional affixes depending on the style, register and genre of texts;
- study of the frequency of case affixes of nouns of different semantic classes;
- comparative study of the frequency of case affixes on data of corpora of Turkic languages;
- extending the list of studied affixes including verbal ones (tense affixes, participle affixes, converb affixes, etc.);
- development of ways of visualizing the distribution of Tatar and Turkic inflectional affixes.

Acknowledgements

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

REFERENCES

1. International Quantitative Linguistics Association. URL: http://www.iqla.org.

2. Kopotev M.V. *Towards building the frequency of Russian grammar: the case system on corpus data* [K postroeniyu chastotnoy grammatiki russkogo yazyka] In *Slavica Helsingiensia.* – 2008. – Vol. 34. Pp. 136–151.

3. Mačutek, J., & Čech, R. (2012). *Frequency and Declensional Morphology of Czech Nouns.* In *The international quantitative linguistics conference QUALICO.* Pp. 26-29.

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

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

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

7. Sak H., Güngör T., and Saraçlar M. (2008). Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus. In *Advances in Natural Language Processing, Lecture Notes in Computer Science.* Volume 5221, 2008. Pp. 417–427.

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

9. *Tatar Grammar* [Tatarskaya grammatika]: in 3 volumes. Kazan: Tatar publishing company, 1993. V. 2: Morphology. 397 p. (In Russian).

10. *Tatar Grammar* [Tatar grammatikası]: in 3 volumes. Moscow: Insan, Kazan: Fiker, 2002. V. 2. – 448 p. (In Tatar).

11. Corpus of Written Tatar. URL: http://corpus.tatfolk.ru/index_en.php.

12. Tatar National Corpus. URL: http://corpus.antat.ru.

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

14. Guzev V.G., Burykin A.A. (2007). *Common Structural features of agglutinative languages* [Obshchiye stroevye osobennocti agglutinativnykh yazykov]. In *Acta linguistica Petropolitana. Proceedings of ILR RAS. V. 3:* 1. Pp. 109–117.

15. Becerra-Bonache L., Jimenes-Lopez M.D. (2015). A grammatical inference model for measuring language complexity. In *Advances in Computational Intelligence – 13th International Work-Conference on Artificial Neural Networks,* IWANN 2015, Palma de Mallorca, Spain, June 10–12, 2015. Proceedings, Part I (2015). Pp. 4–17.

16. Berta A. (1998). Tatar and Bashkir. In Johanson, Lars & Csató Éva Agnes (ed.). 1998. *The Turkic languages*. London: Routledge. pp. 283–300.

17. *The Turkic Languages*. L. Johanson & É.Á. Csató (eds). Routledge (1998).

18. Haspelmath M. An empirical test of the Agglutination Hypothesis. In *Universals of Language Today. Studies in Natural Language and Linguistic Theory Volume* 76, 2009, pp. 13–29.

19. Juola, P. 2008. Assessing Linguistic Complexity. In Miestamo, Matti, Kaius Sinnemaki and Fred Karlsson (eds.). *Language Complexity: Typology, Contact, Change.* Amsterdam: John Benjamins Press.

20. Kusters, W. (2003). *Linguistic complexity*. Netherlands Graduate School of Linguistics.

21. Menges K. H. *The Turkic Languages and Peoples: An Introduction to Turkic Studies. Wiesbaden*: Harrassowitz, 1994.

MODELING OF TENSE SYSTEM IN AGGLUTINATIVE LANGUAGES WITH SEMANTIC SITUATIONS

Zhandos Zhumanov

al-Farabi Kazakh National University,
71 al-Farabi Ave., Almaty 050040, Kazakhstan
z.zhake@gmail.com

The paper provides a formal description of verbs tenses in agglutinative languages using semantic situations and regular expressions on example of Kazakh language. We describe features of Kazakh language's tense system, semantic situations and their representation in the form of regular expressions. Kazakh language's tense system is simulated with semantic situations. A practical use example of proposed solution is provided. There are practical results.

1. Introduction

Tense category is one of the most important grammatical categories of natural languages. In various languages, this category has a different grammatical representation, but its purpose is always the same, that is to show how a text relates to a timeline, adopted in the language. Except for the obvious way to reference time (past, present, future), there are several expressions of time, which in some languages are not obvious. Formal description of tense category is complicated by the fact that the same time moment can be expressed using different grammatical structures. This results in the fact that in natural language processing it can be difficult to analyze time expressions.

To solve this problem, we can use the fact that expression of time in natural languages consists of two components: grammatical and semantic. And the second component, in our opinion, is more important. For example, in machine translation very often tense of the same sentence in different languages may not match.

2. Tense category in linguistics

According to (Fabricius-Hansen, 2006), in grammar tense is a category that determines the position of the action on the timeline. Tense category uses an indication of time tied to some moment. Referencing time to the present moment is sometimes called absolute time, referencing to the moment other than current is called relative time. The first case

gives us so-called "simple" temporary structures (past simple in English, жедел өткен шақ in Kazakh). Using relative time can be seen in such grammatical constructions as "future in the future" or "future in the past" (i.e. perfect tenses in English). Relative time can show precedence of some events with respect to others, their simultaneity or their order.

Time is a grammatical category of the verb. Consequently, grammatical transformations necessary for expression of described senses are applied to verbs. Synthetic and analytical verbs are used for this purpose (Ярцева, 1990). In synthetic form several morphemes are combined in a single word (eg, "reads", "writing"). In analytical form primary and secondary meanings of the word are expressed separately (for example, "I will read", "will write").

3. Description of tense system in Kazakh language

7 tenses can be identified in Kazakh language. Three of them are past tense, one is present, two are future and one can express both present and future (Мурзагельдинова, 2009). All of them are presented in Table 1.

Table 1

Tense system in Kazakh language

Tense name	Construction	Meaning	Example
Past tense			
Жедел өткен шақ	Base verb + ды/ді/ты/ті + personal endings	Recent past	Мен барғанмын. I have gone.
Бұрыңғы өткен шақ (3 ways of construction)	Base verb + қан/кен/ған/ген + personal endings	Long gone past	Мен барыппын. I went.
	Base verb + ып/іп/п + personal endings		Сен келіпсін. You came.
	Base verb + қан/кен/ған/ген еді/екен + personal endings		Сен кеткен екенсің. You went away.
Ауспалы өткен шақ	Base verb + атын/етін/йтын/йтін + personal endings	Long gone past, sometimes relative past	Мен бұрын хат жазатынмын. I wrote a letter long ago.

Present tense			
Нақ осы шақ	Отыр, түр, жүр, жатыр + personal endings	Continuing present	Мен жатырмын. I am lying.
	Base verb + ып/іп/а/е/и auxiliary verb + personal endings		Сен ойлап жүрсің. You are thinking
Ауспалы осы шақ	Base verb + а/е/й + personal endings	Present or near future	Мен ойнаймын. I play. or I will play.
Future			
Болжалды келер шақ	Base verb + ар/ер/р + personal endings	Presumable future	Мен келермін. I might come.
Мақсатты келер шақ	Base verb + мақ/мек/бақ/бек/пақ/пек/шы/ші + personal endings	Future with intent	Сен жазбақсын. You are going to write.

As you can see from the table tenses in Kazakh are mainly constructed using synthetic forms of the verb. One form of Бұрыңғы өткен шақ and Нақ осы шақ use analytical form of construction.

4. Description of semantic situations

One of the key components in each natural language is the meaning of the phrases and expressions. Regardless of the language expressions with similar meanings are used in similar contexts. But often those expressions have different grammatical structures in different languages. Examples of such contexts may be “greeting”, “description of relationships”, “use of ordinal numbers”, etc. It is possible to create a set of semantic “situations” that have the same meaning in all languages.

Formally semantic situations can be described as follows:

<language> ::= <sentence> | <language><sentence>

<sentence> ::= <set expression>

<sentence> ::= <word-combination> | <word-combination><sentence>

<word-combination> ::= <word’s grammatical form> | <word-combination><word’s grammatical form>

<word-combination> ::= <word> | <word-combination><word>
<word's grammatical form> ::= <word><grammatical features>
<word's grammatical form> ::= <grammatical features><word>
<grammatical features> ::= <affix> | <preposition> | <...>
<word> ::= <noun> | <adjective> | <...>
<semantic situation> ::= <set expression> | <word-combination> | <word's grammatical form>

Conditions of certain verb tenses' usage can also be a semantic situations that have a meaning (sense) and are expressed in synthetic or analytical form. Language's tense system can be represented as a set of semantic situations. But to do this in addition to a verbal description of these situations we need a writing format that supports computer processing of the resulting set.

5. Regular expressions and their use in natural language processing

In text oriented software technologies there is a tool that significantly simplifies processing of text data called regular expressions. Regular expressions can be thought of as a mini programming language, which has one specific purpose: to find a substring in large string expressions. It is not a new technology; initially it appeared in UNIX environment and is commonly used in Perl programming language. However, each developed programming language has its own implementation of regular expressions, in general corresponding to Perl regular expressions (Фридл, 2003).

Semantic situations that use a synthetic form of construction can be written as regular expressions. Writing regular expressions for semantic situations with analytical form of construction may have some difficulties if relevant grammatical rules have no patterns (i.e. exceptions).

6. Simulation of tense system using semantic situations

Using semantic situations and writing them in the form of regular expressions as described above we can describe verb tenses in agglutinative languages. This article uses regular expression syntax of C# programming language. In other programming languages there may be minor differences.

Table 2

Tense system of Kazakh language in the form of semantic situations

Жедел өткен шақ	@»\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(ді|ды|ті|ты)(м|н|ныз|ніз|к|қ|ндар|ндер|ныздар|ніздер)?\b»
Бұрынғы өткен шақ 1	@»\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(ып|іп|п)(пын|пін|сын|сін|сыз|сіз|ты|ті|пыз|піз|сындар|сіндер|сыздар|сіздер)\b»
Бұрынғы өткен шақ 2	@»\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(ған|ген|қан|кен)(мын|мін|сын|сін|сыз|сіз|мыз|міз|сындар|сіндер|сыздар|сіздер)?\b»
Бұрынғы өткен шақ 3	@»\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(ған|ген|қан|кен) екен(мін|сін|сіз|біз|сіндер|сіздер)?\b”
Ауыспалы өткен шақ	@»\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(атын|етін|йтін|ытын) еді(м|н|ніз|к|ндер|ніздер)?\b”
Нақ осы шақ	@»\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(ып|іп|п|а|е|я) (отыр|тұр|жүр|жатыр)(мын|мін|сын|сін|сыз|сіз|мыз|міз|сындар|сіндер|сыздар|сіздер)?\b”
Ауыспалы осы шақ	@»\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(а|е|й)(мын|мін|сын|сін|сыз|сіз|ды|ді|мыз|міз|сындар|сіндер|сыздар|сіздер)?\b”
Болжалды келер шақ	@»\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(ар|ер)(мын|мін|сын|сін|сыз|сіз|мыз|міз|сындар|сіндер|сыздар| сіздер)?\b”
Мақсатты келер шақ	@»\b[аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+(мақ|мек|бақ|бек|пақ|пек)(пын|пін|сын|сін|сыз|сіз|пыз|піз|сындар|сіндер|сыздар|сіздер)?\b»

Each one of the expressions contains an element [аәбвгғдеёжзийкқлмнңоөпрстуұүфхһцчшщъыіьэюя]+. Literally it represents an arbitrary sequence of characters in square brackets with length of 1 or more character. Word stems (base verbs) are used in grammatical forms of tenses in the natural languages. For this reason, we have to add an additional requirement to this element: a sequence of characters must be a verb of the natural language.

7. Practical implementation

Proposed representation of tense system in the form of semantic situations can be used to in programming tasks associated with natural languages. We can give an example of machine translation program that translates tense forms from Kazakh language into corresponding forms of English language.

The code bellow searches verb tense forms (semantic situations described in the previous section) in a text. Variable regex contains a description of a semantic situation as a regular expression. The variable text contains analyzed text. Operation regex.Match() performs analysis of matching the text to the regular expression. If correspondence is found, then the analyzed text contains a form of verbal tense. The result of the code can be seen in Fig. 1.

```
for (i = 0; i <this.verb_situation_patterns_1.
Length/3; i++)
{
  Regex regex = new Regex(this.verb_situation_
  patterns_1[i, 1]);
  Match match = regex.Match(text);
  if ((match.Success) && (match.Length == text.
  Length))
    return new Verb_Situation(this.verb_situa-
    tion_patterns_1[i, 0],
        text, this.verb_situation_patterns_1[i,
        2]);
}
```

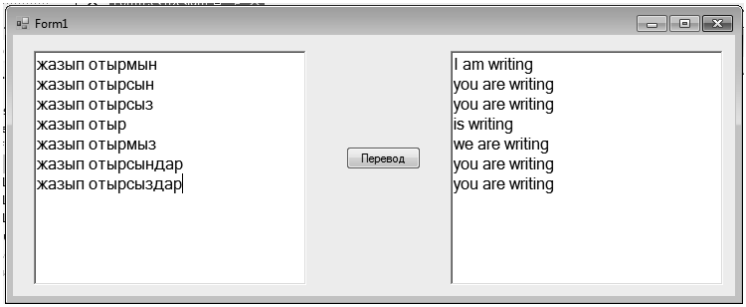


Fig. 1. The result of the analysis and translation of verb forms

As it can be seen from the Fig. 1 the situation “нақ осы шақ” was found correctly and corresponding form in English (present continuous tense) was generated. All possible variations of нақ осы шақ for person and number were found and recognized by the regular expression.

8. Conclusion

This paper is devoted to simulating linguistic tense category. To solve this problem, we propose to use two components of time expressions in natural languages: grammatical and semantic. Semantic component is represented by a set of semantic situations, grammatical – corresponding set of regular expressions.

The paper describes the features of tense system in Kazakh language as an example of agglutinative languages, presents semantic situations and their writing in the form of regular expressions that simulate tense system of Kazakh language. An example of practical use of the proposed solution is provided.

Using of proposed regular expressions method in natural language processing can reduce the amount of code to be written.

Further improvement of proposed solution is related to further development of semantic descriptions of situations in order to enable them better reflect the characteristics of grammatical structures with analytical form of construction.

REFERENCES

Fabricius-Hansen, C. (2006). *Tense. Encyclopedia of Language and Linguistics* (2nd ed.). (pp. 566–573). Amsterdam: Elsevier.

Ярцева, В.Н. (1990). (Editor) *Лингвистический энциклопедический словарь*. Москва: «Советская энциклопедия».

Мурзагельдинова О.И. (2009). *Грамматический справочник-шпаргалка по казахскому языку*. Келешек-2030.

Фридл Дж. (2003). *Регулярные выражения*. Издательство «Питер», 460 с.

COMPUTER –MATHEMATICAL MODELING OF NATIONAL SPECIFICITY OF SPATIAL MODELS IN KYRGYZ LANGUAGE

Sonunbubu Karabaeva, Polina Dolmatova, Aisuluu Imanalieva

sonun2008@mail.ru, Bishkek, Kyrgyzstan, 720020
dolpol@list.ru, Bishkek, Kyrgyzstan, 720040
lunnayakrasa@mail.ru, Bishkek, Kyrgyzstan, 720082

This article describes computer and mathematical modeling of the Kyrgyz language. It is shown that the theory of computer modeling in the study of spatial knowledge using the concepts of natural language (morphological approach), and in the mathematical modeling can be considered as actions with objects are mathematically represented by connected sets. We describe the spatial characteristics of features. The article describes the linguistics spatial relationships of the Kyrgyz language. The article describes the spatial information of the situation. A mathematical model of information interaction is given in this article.

In this paper we propose presentation of one kind of specific words in Kyrgyz language.

Introduction

As research in the field of Natural Language Processing in Turkic world there is a growing need for developing language resources.

Nowadays computer equipment is commonly used in studying foreign languages. To built software for studying and testing natural languages, we began to combine the following ideas: studying language without intermediate language; constructing of random assignments; availability for philologists who may develop exams and electronic study books without the help of software developers.

The computer-based system “Language Education System” (LES) that joins these ideas was developed in [1]. LES is based on the independent computer representation of natural languages’ notions [2], [3].

We implemented some verbs, nouns, adjectives.

In [4] authors proposed representation of complex notions. For that necessary entities (such that 1D, 2D, 3D spaces, motion, animation, avatar, transformation etc.) should be used.

1. Main definitions and hypotheses

In [5] it was proposed to develop independent computer representation of natural languages on the base of the following

Definition 1. If low energetic outer influences can cause sufficiently various reactions and changing of the inner state of the object then it is said to be an affectable object, or a subject, and such outer influences are said to be commands (these reactions and changing are implemented by means of inner energy of the object or of outer energy entering into object besides of commands).

Remark 1. In programming languages, statements are divided into declarations and commands usually. But all declarations also may be considered as commands. Also, narrative and interrogative sentences in natural languages also are implicit imperative ones.

Definition 2. A system of commands such that any subject can achieve desired sufficiently various consequences from other one is said to be a language.

These definitions unite humans and computers and their languages.

Remark 2. The other subject may be the same. For instance, a person being afraid of forgetting writes down an instruction or a list for her/himself to use it in future.

Hypothesis 1. A human's genuine understanding of a text in a natural language can be elucidated by means of observing the human's actions in real situations corresponding to this text.

Definition 3. Let any "notion" (word of a language) be given. If an algorithm acting at a computer: performs (generated randomly) sufficiently large amount of situations covering all essential aspects of the "notion" to the user; gives a command involving this "notion" in each situation; perceives the user's actions and performs their results clearly on a display; detects whether a result fits the command, then such algorithm is said to be a computer interactive presentation of the "notion".

In general, the environment for the user consists of (generated randomly) constant objects, moving-transforming objects and controlled objects. Random generation of auxiliary objects and their positions is necessary to distinguish the "notion" among other words being used in the command and arising circumstances which are not sufficient for the "notion".

Remark 3. Certainly, commands are to contain other words too. But these words must not give any definitions or explanations of the "notion".

Definition 4. If all words being used in Definition 3 can be unknown for the user nevertheless s/he would be able to fulfil the meant action (because it is the only natural one in this situation) then the notion (word of a language) is said to be primary. If any words known to the user are necessary then the notion is said to be secondary.

Thus, a natural hierarchy of notions arises.

By our experience, verbs PUT, TAKE, PUSH, COMPLETE, ENTER, GO OUT, OPEN, CLOSE, SYMMETRIZE and simple nouns such as BALL, STICK, BOX are primary.

Remark 4. By such a way not only notions presenting real objects and relations but notions presenting imaginary objects can be presented; we demonstrated natural interactive performing of abstract spaces (Borubaev A.A., Pankov P.S., 1998). By the methods proposed in this paper, numerous computer games can be developed to "adequate" presenting of notions corresponding to fictional and fabulous objects.

Remark 5. Modern displays are formally discrete but they are perceived as continuous. So, continuous motion and transformations can be implemented.

Usually a notion is defined as a mental image or representation. But there is a more definite hypothesis in pattern recognition: various images corresponding to a same notion form a "compact" set (the term "compact" is understood informally). In [6] it was proposed

Hypothesis 2. A child or a human learning a natural language without references to known languages hearing a notion in various situations begins to form a kind of mathematical model in mind corresponding to this notion by means of trial and error method and attempts to fulfil operations similar to mathematical ones: closing and compactification. After successful completing such operations, the human feels "mastering" this notion.

Hypothesis 3. Any notion has the minimal model (involving minimal number of entities in Occam's sense). These entities can be mathematical involved into such verbs as MOVE – geometry, CUT – topol-

ogy, FLEX, TIE – 3D, physical (DROP, BOIL, FREEZE), chemical (BURN), affectable (GIVE, SHOW, LISTEN, HEAR), etc. We use the term "affectable" instead of "animated" because it is related to computers as well as to humans and animals in present-day speech.

Remark 6. From this point of view, chemical essences unite with some physical ones into "sufficiently transforming" ones.

Hypothesis 4. Up-to-date multimedia computer equipment is sufficient to model situations necessary to teach and detect genuine understanding of vital notions in natural languages.

2. Peculiarities of spatial relations in Kyrgyz language

Peculiarity of Kyrgyz language is that parts of space related to any object from the viewpoint of the subject and taking gravitation and direction of observation and motion in account are presented as nouns, with corresponding cases: Dative, Locative, Ablative, Accusative.

YCT(Y) – upper-space, АСТ(Ы) – before-and-lower-(observed)-space, ИЧ – interior, СЫРТ – exterior, ЧЕК – boundary-strip, СОЛ – left-space, ОҢ – right-space, ОРТО – middle-spot, ЖАКЫН – near-space, ЖАН (ЫНДА) – near-space, АРА – between-space, АЛД (Ы) – before-forward-space, АРТ(Ы) – behind-space, КАРШЫ– opposite- space, МАНДАЙ – in front of- space.

For example:

The lamp is over the chair – (adverb)

ЧЫРАК ОТУРГУЧТУН YCTYНДӨ – (literatim) (In Kyrgyz language the word "YCTYНДӨ" is adverb).

The lamp (is) in the chair's upper-space.

Who is higher of the director? – upper-space. (In English language the word "HIGHER" is adverb. Somebody is above director – the meaning)

ДИРЕКТОРДУН YCTYНДӨ КИМ БАР? – (conversational) (In Kyrgyz language the word "YCTYНДӨ" is Pronoun).

3. Mathematical and computer models' definitions

Definitions of mathematical and computer models of natural language's notion were developed to implement the independent representation of natural languages [7]. Sets theory was chosen as a mathematical

apparatus because the natural languages' notions can be considered as the actions with objects mathematically represented by connected sets.

Definition 5. Mathematical model of natural language's notion is:

– the list of sets, some of which are the subsets of other sets from this list, with the groups of sets for random choice;

– the list of possible motions and other transformations of these sets;

– the lists of allowable and unallowable relations between sets during their transformations;

– the list of necessary relations (intersection, embedding) in the time sequence;

– the list of assignments including the notion,

that give its sufficiently complete and adequate representation.

Definition 6. Computer model of natural language's notion is the model based on the mathematical model of this notion with graphical representation of sets.

Example 1. Mathematical model of the noun СОЛ.

The media where the notion is demonstrated is the two-dimensional geometric space.

Sets: K is the cursor, three sets of objects: $L=\{P_i \mid i=1..n\}$, $R=\{P_i \mid i=1..m\}$, $C=\{P_i \mid i=1..l\}$, $L\cap C=\varnothing$, $L\cap R=\varnothing$, $C\cap R=\varnothing$.

Assignment: «P_{i_0} P_{j_0}-НЫН СОЛУНДА ЖАТАТ».

Unallowable relations:

$$\exists j=\overline{1,m}:(K\in P_j)\wedge(P_j\in C);\ \exists i=\overline{1,n}:(K\in P_i)\wedge(y(P_i)\neq y(P_{j_1})).$$

Necessary relations for moments $t_0 < t_1$:

$$t_0:(\forall i=\overline{1,n}\ \forall j=\overline{1,m}\ \forall k=\overline{1,l}\ \forall P_i\in L\ \forall P_j\in C\ \forall P_k\in R$$
$$x(P_i)<x(P_j)<x(P_k))\wedge(y(P_{i_0})=y(P_{j_0}));$$
$$t_1:\ (K\in P_{i_1})\wedge(P_{i_1}\in L)\wedge(y(P_{i_1})=y(P_{j_1})).$$

4. Computer presentation

The following program can be built to demonstrate moving a little object by a verbal order, for example:

1) The user observes the media and inputs the text:

"КЫЗЫЛ ТОПТУ ОТУРГУЧТУН АСТЫНА КОЙ!"

(Put the red ball under the chair!)

2) The computer analyses (there are corresponding algorithms in Kyrgyz language):

КЫЗЫЛ ТОП+НЫ ОТУРГУЧ+НЫН АСТ+Ы+НА КОЙ!

3) The computer detects affixes:

КЫЗЫЛ ТОП+Accusative-case-affix ОТУРГУЧ+Genitive-case-affix АСТ+Possessive-affix+Dative-case-affix КОЙ!

4) If these analysis or detection are failed or there is not such sentence-pattern then

the computer responds:

"ТУУРА ЭМЕС БУЙРУК" ("Wrong command")

otherwise

5) The computer searches objects corresponding to revealed stems of words in the media. If "Red Ball" or "Chair" do not exist then it responds

"МЫНДАЙ НЕРСЕ(ЛЕР) ЖОК" ("Absence of such objects")

otherwise

6) The computer moves the object to the spot slowly.

7) The computer finishes the action announcing:

"КЫЗЫЛ ТОП ОТУРГУЧТУН АСТЫНДА" ("The red ball is under the chair").

Such computer programs will be developed for demonstration spatial relations in Kyrgyz language.

Conclusion

We hope that mathematical modeling approach would be useful for profound investigation of Kyrgyz language. Computer systems that are based on the mathematical and computer models of notions can be used for teaching Kyrgyz language. Suggested construction of spatial notions' models is a significant part of independent computer presentation of Kyrgyz language.

REFERENCES

1. Pankov P., Dolmatova P. (2009) Software for Complex Examination on Natural Languages. *Human Language Technologies as a Challenge for Computer Science and Linguistics: Proceedings of the 4th Language and Technology Conference*. – Poznan, Poland, pp. 502–506.

2. Pankov P., Dolmatova P. (2008) Algorithmic Language and Classification of Verbs for Computer-Based Presentation. *Academic Review*, No. 1 (7). – American University of Central Asia, Bishkek, pp. 233–239.

3. Bayachorova B.J., Pankov P.S. (2009) Independent Computer Presentation of a Natural Language. In: *Varia Informatica*. Polish Information Processing Society, Lublin, pp. 73–84.

4. Pankov P., Dolmatova P. (2009) Algorithmical Language for Computer-Based Presentation of Notions. *IKECCO'2007: Proceedings of the 4th International Conference on Electronics and Computer*. – Suleyman Demirel University, Almaty, pp. 274–279.

5. Pankov P., Bayachorova B., Juraev M. (2012) Mathematical Models for Independent Computer Presentation of Turkic Languages. *TWMS Journal of Pure and Applied Mathematics,* Volume 3, No.1, pp. 92–102.

6. Pankov P., Bayachorova B., Juraev M. (2014) Occam's razor in mathematical models of notions of Turkic languages. *Abstracts of V Congress of the Turkic World Mathematicians (Kyrgyzstan, Bulan-Sogottu, 5–7 June, 2014)* / Ed. A.Borubaev. Bishkek: Kyrgyz Mathematical Society. – P. 310.

7. Dolmatova P. (2014) Mathematical Modeling of Notions as the Basis of Independent Representation of Natural Languages. *Abstracts of V Congress of the Turkic World Mathematicians* (Kyrgyzstan, Bulan-Sogottu, 5–7 June, 2014) / Ed. A.Borubaev. Bishkek: Kyrgyz Mathematical Society. – P. 302.

HOW MANY MODALITIES OF CASE ASSIGNMENT ARE THERE IN TATAR?

Ekaterina Lyutikova[a,b], Dilya Ibatullina[b]

[a] Lomonosov Moscow State University, Leninskiye Gory, MSU, Moscow 119991, Russia
[b] Sholokhov Moscow State University for Humanities, 16–18 Verkhnyaya Radishchevskaya, Moscow 109240, Russia
[1]lyutikova2008@gmail.com

The paper discusses formal models of structural case assignment and their applicability to Tatar data. We survey two approaches — Chomskyan model of case assignment by functional heads and Marantzian model of configurational case marking — and their application to Sakha, which received much attention in the recent literature. Addressing linguistic evidence from Tatar, we identify a number of contexts where differential case marking is observed. We claim that neither the purely configurational model nor the hybrid model combining Chomskyan and Marantzian modalities of case assignment can account for Tatar data. We argue that the Chomskyan model of case assignment under AGREE can be adjusted in order to successfully account for the full range of Tatar case phenomena. Specifically, we distinguish between different "nominative" forms in Tatar. We suppose that nominative direct objects and nominative possessors are in fact caseless structurally deficient nominals that do not need abstract case in order to be licensed. Nominative subjects, on the other hand, are case-marked and control agreement on the predicative functional head. The nominative form of the prepositional phrase complement is neither a caseless form (as it is available for ezafe-3 nominals) nor a subject case. Since the distribution of genitive and nominative forms governed by prepositions is purely morphological, we propose that they are morphological exponents of an abstract "postpositional" case.

1. Introduction

Recently, significant advances have been made in the investigation of case systems across languages, including typologically-oriented studies and formal modelling of Case. The most controversial issue regarding case is an apparent lack of a consistent interpretation of this grammatical category, both within an individual case system as well as crosslinguistically. This is especially true of the so-called grammatical, or structural, case (Kurilovich, 1962; Chomsky, 1986). Comparing structural case with other grammatical categories defined in the

nominal or the verbal domain, we may wonder what semantics the case morpheme bears except for a purely configurational information. Thus, we may ask, for example, what contribution to the semantic interpretation Russian accusative case morpheme makes, aside from signaling the syntactic dependence of the noun phrase (NP) on a certain verb or preposition. At the same time, we may ask whether Russian and Tatar accusative shall be analyzed along the same lines, keeping in mind that although they both mark the direct object, they differ significantly in their distribution.

Due to this primarily syntactic function of case, existing theoretical studies of case mostly focused on the syntactic contexts and mechanisms of case assignment, rather than on the semantic or pragmatic characteristics of a noun phrase underlying the choice of a given case morpheme. This common property of different formal analyses of case assignment is easy to observe in their treatment of differential object marking (DOM, see Bossong, 1985; Aissen, 2003). Interpretational effects accompanying the use of a specific case affix, like telicity of (1a) and atelicity of (1b) from Estonian, or animacy in (2a) and inanimacy in (2b) from Ossetic, or specificity in (3a) and non-specificity in (3b) from Turkish, are usually considered as a by-product of the characteristics of noun phrases that condition the case assignment.

(1) a. Ta ehitas silla.
he build.PST.3SG bridge.GEN
'He built a/the bridge'.
b. Ta ehitas silda.
he build.PST.3SG bridge.PART
'He was building a/the bridge'.

(2) a. æž axuırgænæǯ-ı fedton.
I teacher-GEN see.PST.1SG
'I saw a/the teacher'.
b. æž činıg fedton.
I book see.PST.1SG
'I saw a/the book'.

(3) a. dün çok garip kitap-lar-ı oku-du-m.
yesterday very strange book-PL-ACC read-PST-1SG
'Yesterday I read the very strange books'. (Kornfilt, 1997: (1019))

b. dün çok garip kitap-lar oku-du-m.
yesterday very strange book-PL read-PST-1SG
'Yesterday I read very strange books'. (Kornfilt, 1997: (1018))

In this paper, we aim to show that the mere existence of case marking variation, as well as the factor(s) underlying this variation shall be taken into account when the mechanisms of case assignment are modelled. Our contribution is based on the evidence from Tatar. We identify syntactic contexts of case marking variation and the parameters triggering the choice of a case morpheme, and argue that these data call for a multi-modal mechanism of case assignment.

The paper is organized as follows. In Section 2, the two competing models of structural case assignment are discussed. Section 3 represents different accounts of case marking in another Turkic language, Sakha, which received much attention in recent literature. In Section 4, we examine Tatar data and show that if the case marking variation is taken into account, we need many modalities of case assignment. Section 5 sketches the analysis which allows us to deal with case marking in Tatar in a uniform way.

2. Formal models of case assignment

In the formal syntactic literature, two major ideas can be found about how the case morphology ends up appearing on a given noun phrase. The first approach, which is the most widely assumed approach of the generative grammar, treats case as a primarily syntactic phenomenon that licenses NPs; the second approach, put forward in the work of Alec Marantz, considers case as a postsyntactic, purely morphological phenomenon.

The modern Minimalist syntactic approach to case is a development and elaboration of the Case theory of the Government and Binding model (Chomsky, 1981, 1986). Case is considered as an unvalued uninterpretable feature of a noun phrase that has to be valued in order to prevent the derivation from crashing. In the Chomsky-style model, Case is assigned to a noun phrase under AGREE relation with a dedicated case-assigning head. There are two kinds of case-assigning heads: lexical heads, that assign case to their own arguments exclusively, and functional heads, that assign case to the nearest goal NP available in

their c-command domain. The case assigned by a lexical head is called inherent, the case associated with a functional head is called structural case (Chomsky, 2000, 2001).

The characteristic properties of a structural case are: (i) its independence from a semantic role; (ii) its somewhat non-local nature and (iii) the non-obligatoriness of its realization. In view of these properties, three major structural cases are usually recognized: nominative, assigned by the finite predicative head T; accusative, assigned by the transitive light verb head *v*; and genitive, assigned by the (possessive) determiner head D.

A competing morphological case approach dates back to (Marantz, 1991) and is further elaborated in (Bittner, Hale, 1996), (McFadden, 2004), (Bobaljik, 2008). The basic idea behind this approach is that case is assigned to noun phrases on a configurational basis, depending on the presence of other noun phrases ("case competitors") in the same local domain.

Marantz distinguishes four distinct kinds of case, forming a disjunctive Case realization hierarchy (Marantz, 1991: 24):

- Lexically governed case (i.e., case determined by the lexical properties of a particular item, such as quirky-case-assigning verbs in Icelandic, or adpositions in many languages)
- "Dependent" case (accusative case and ergative case)
- Unmarked case (e.g., nominative case assigned to any NP in a clause; genitive case assigned to any NP inside an NP/DP)
- Default case (assigned to any NP not otherwise marked for case).

This Case realization disjunctive hierarchy determines the order in which the different kinds of case shall be assigned. First, lexically-governed case (analogous to the inherent case described above) is assigned. Next, the rule of dependent case assignment applies; it requires a configuration where there are at least two caseless NPs in the clausal domain. If this requirement is met, the lower NP is marked with the "depended" accusative case in accusative languages, or the higher NP is marked with "depended" ergative case in ergative languages. Then, it is the turn of the unmarked case rule that marks any still case-less NP in a given syntactic domain with the dedicated case. Finally, if neither of the previous rules applied to an NP, it receives the default case.

The basic innovations in the Marantz-style system include a more elaborate definition of case competition domains, as well as on different modes of dependent case assignment. Thus, it has been proposed that not only the higher or the lower NP can receive case in a "depended mode" – it is possible that in some tripartite case systems like the one in Nez Perce both the higher and the lower NPs are marked. Within the clause, more domains for case competition have been distinguished, e.g. VP and CP, which allowed to subsume dative case assigned in ditransitive constructions under a similar analysis.

It should be noted that although morphological case assignment is construed as independent from agreement of lexical or functional heads (i.e. AGREE operation), the morphological case marking can in principle feed into the agreement process. Thus, J. Bobaljik (Bobaljik, 2008) reinterprets E. Moravcsik's (Moravcsik, 1974, 1978) hierarchy as the hierarchy of accessibility of case-marked NPs as controllers of agreement:

(4) unmarked case ≫ dependent case ≫ lexical/oblique case

The two approaches to case assignment are mostly considered as incompatible. However, a recent paper by M. Baker and N. Vinokurova (Baker, Vinokurova, 2010) claims that the two approaches represent different modalities of case assignment that can coexist in the grammar of one and the same language, Sakha being one example. In the next section, we discuss Baker and Vinokurova's analysis in more detail.

3. Case assignment in Sakha

(Baker, Vinokurova, 2010) claim that the distribution of morphological case in Sakha requires a hybrid approach to the assignment of case. Specifically, the authors argue that the distribution of accusative and dative case in Sakha can only be accounted for configurationally, whereas nominative and genitive case require an account in terms of case assignment by functional heads.

The two major parameters that distinguish between Marantz-style and Chomsky-style modalities of case assignment are (i) whether case assignment is sensitive to the presence of another noun phrase (NP) or to the presence of a functional head (F) in the local domain, and (ii) whether case assignment is followed by the obligatory agreement with the case assigner. The parameters' valuation for Sakha four cases is represented in Table 1.

Table 1

Modalities of case assignment in Sakha

Cases:	Nominative	Genitive	Dative	Accusative
NP/F	F (finite T)	F (determiner)	NP (verbal domain)	NP (clausal domain)
Agreement	+ (predicative)	+ (possessive)	–	–
Modality of case assignment	Chomsky-style	Chomsky-style	Marantz-style	Marantz-style

Baker and Vinokurova propose that nominative is assigned to the subject by the agreeing functional head (5); interestingly, only one agreement marker per nominative subject is available (6).

(5) Masha-qa **at-tar** ber-ilin-ni-**ler** || *ber-ilin-ne.
Masha-DAT horse-PL give-PASS-PST-3PL || give-PASS-PST.3SG
'The horses were given to Masha.' (Baker, Vinokurova, 2010: (66))

(6) a. **En** süüj-büt e-bik-**kin**.
you win-PTPL AUX-PTPL-2SG
b. **En** süüj-bük-**kün** e-bit.
you win-PTPL-2SG AUX-PTPL
c. *En süüj-büt e-bit.
you win-PTPL AUX-PTPL
d. *En süüj-bük-kün e-bik-kin.
you win-PTPL-2SG AUX-PTPL-2SG
'The result is that you won'. (Baker, Vinokurova, 2010: (76))

In a similar vein, the genitive possessor is only available when the head noun bears the possessive agreement suffix; Baker and Vinokurova suppose that Sakha noun phrase only licenses the genitive possessor when it is headed by an agreeing functional head D. Again, multiple agreement with the case-marked NP is prohibited.

So, nominative and genitive case assignment in Sakha is in one-to-one correspondence with the availability of an agreeing functional head. This modality of case assignment contrasts radically with the conditions on the assignment of dative and accusative cases. First of all,

dative and accusative case assignment is never dependent on agreement with whatever functional head; dative- or accusative-marked NPs can only control agreement in very particular contexts where the genitive possessor or the nominative subject has subsequently raised and acquired dative or accusative, respectively:

(7) a. **Misha-qa** beqehee at-**a** öl-lö.
Misha-DAT yesterday horse-**3SG** die-PST.3SG
'Misha's horse died on him yesterday.' (Baker, Vinokurova, 2010: (46a))

b. Min **ehigi-ni** bügün kyaj-yax-**xyt** dien erem-mit-im.
I you-ACC today win-FUT-2PL that hope-PST-1 SG
'I hoped you would win today.' (Baker, Vinokurova, 2010: (37a))

Secondly, dative and accusative case assignment is sensitive to the presence of another NP in the specific local domain. Dative is assigned to the higher NP in the verbal domain (VP) if there is another caseless NP within the same domain. So the configurational dative found in ditransitives (8a) and causatives of transitives (8b) is the "dependent" case in the verbal domain.

(8) a. Min [Masha-**qa** **kinige** bier]-di-m.
I [$_{VP}$ Masha-DAT book give]-PST-1SG
'I gave Masha books/a book.' (Baker, Vinokurova, 2010: (11a))

b. Misha [Masha-**qa** **miin** sie-t]-te.
Misha [$_{VP}$ Masha-DAT soup eat-CAUS]-PST.3SG
'Misha made Masha eat soup.' (Baker, Vinokurova, 2010: (21b))

Accusative case, according to Baker and Vinokurova, is a "dependent" case, too; it is assigned in the clausal domain to the lower of the two caseless NPs. This analysis captures perfectly the important property of the Sakha clause syntax: accusative direct objects appear outside the VP (that is, strictly to the left of VP-level adverbials and indirect objects), whereas unmarked, or nominative, direct objects appear within the VP, in their base position. When the definite direct object raises out of VP to avoid existential closure, it lands in the clausal

domain of case competition and is assigned accusative in the presence of a still-caseless subject NP:

(9) **Min** kinige-**ni** [Masha-qa ~~kinige~~ bier]-di-m.
I book-ACC [$_{VP}$ Masha-DAT ~~book~~ give]-PST-1SG
'I gave the book to Masha'. (Baker, Vinokurova, 2010: (11b))

The crucial argument in favor of the Marantz-style accusative case assignment in Sakha comes from the dependent clauses with accusative subject like the one in (7b). It appears that the availability of accusative does not depend on the presence of an accusative-assigning head in the main clause, cf. (10a) with an intransitive verb and (10b) with a passive verb; what really matters is the presence of another NP in the main clause (10c).

(10) a. **Keskil** Aisen-**y** [kel-bet dien]
Keskil Aisen-ACC come-NEG.AOR.3SG that
xomoj-do.
become.sad-PST.3SG
'Keskil became sad that Aisen is not coming.' (Baker, Vinokurova, 2010: (39a))

b. **Sargy** kim-**i** daqany [tönn-üm-üö dien]
Sargy who-ACC PRT return-NEG-FUT.3SG that
erenner-ilin-ne.
promise-PASS-PST.3SG
'Sargy was promised that nobody would return.' (Baker, Vinokurova, 2010: (40))

c. Bügün munnjax-xa Masha-(***ny**) [ehiil
today meeting-DAT Masha-(*ACC) [next.year
Moskva-qa bar-ya dien] cuolkajdan-na.
Moscow-DAT go-FUT.3SG that] become.certain-PST.3SG
'It became clear today at the meeting that Masha'll go to Moscow next year.' (Baker, Vinokurova, 2010: (42a))

Thus, Baker and Vinokurova's analysis of structural case assignment in Sakha makes use of both Chomsky-style and Marantz-style modalities. Recently, some alternatives to this hybrid account of Sakha data have been proposed. T. Levin and O. Preminger (Levin, Preminger, 2015) argue for a purely configurational account of Sakha case

assignment. They propose that genitive and nominative are unmarked cases of the nominal and clausal domain, respectively. Elaborating on the agreement model of (Bobaljik, 2008) and (Preminger, 2011), they suggest that agreement in both T and D in Sakha is case-discriminating in the sense that only noun phrases with unmarked case are visible as goals. Agreement obeys locality, to the effect that some structural configurations can lack an appropriate goal, as in (11). In such cases, agreement fails, but this failure does not yield ungrammaticality, as in Chomsky-style model; instead, a default form of the agreeing head is used, which happens to be zero for D (11a) and coincides with 3SG for T (11b) in Sakha.

(11) a. terilte-ni salaj-aaccy
company-ACC manage-AG.NOML
'the manager of the company' (Levin, Preminger, 2015: (7a))
b. Caakky-ny sorujan ötüje-nen
cup-ACC intentionally hammer-INST
aldjat-ylyn-na.
break-PASS-PST.3SG
'The cup was broken intentionally with a hammer.' (Levin, Preminger, 2015: (13))

In (Kornfilt, 2013), (Kornfilt, Preminger, 2014) a further refinement of the configurational analysis is put forward: "nominative" and "genitive" are simply descriptive labels for caseless NPs that undergo spell-out in the clausal and nominal domain, respectively. In this system, agreement targets caseless NPs exclusively, and only caseless NPs can undergo raising and compete for a "dependent" case.

Interestingly, the obvious question is hardly even noticed in the literature: what motivates the differential case marking available for a noun phrase in a number of structural configurations (direct object, possessor, embedded subject), and how this variation is implemented within each proposed system. It seems that the configurational case assignment model has only one explanatory mechanism: if a noun phrase exhibit differential (structural) case marking, it can belong to different domains of case assignment. Thus, differential direct object marking in (8a) and (9) results from different structural positions of the direct object. Raising of the object NP out of the base position is triggered by interpretational factors, specifically, by the need of the definite noun

phrase to take a wide scope with respect to the VP, which is generally considered as the domain of existential closure.

In the richest system of (Baker, Vinokurova, 2010), there is one more option of analysis of the unmarked noun phrase: it can be (pseudo) incorporated into the verbal (or nominal) stem, and in this way escape the need for case assignment. If, as Kornfilt and Preminger propose, "nominative" is the caseless form of a noun, we are left with only one explanatory idea: case variation means structural position variation.

Chomsky-style case assignment is more flexible in that it allows both NP-external and NP-internal factors triggering case variation to be easily implemented within the model. Factors like telicity, perfectivity or polarity are naturally conceived as (features of) functional heads that assign case to NPs directly or influence the case-assigning abilities of other heads. The impact of factors internal to the noun phrase, like its structural type (noun/pronoun) or its formal and semantic features (animacy), can be analyzed as a split morphological realization of the same syntactic case. Besides, as case assignment obeys at least phase-level locality, the positional alternative is also an option for the Chomsky-style model. Moreover, if we adopt the hypothesis that some structural types of noun phrases, i.e. (pseudo)incorporated nouns and NPs (Massam, 2001, Baker, 2009) or structurally deficient Small Nominals (SNs) – lexical NPs or QPs (Pereltsvaig, 2006, Danon, 2006) do not need (and do not receive) case in order to be licensed, an additional type of case marking variation – between case-marked and caseless noun phrase – emerges (see (Lyutikova, 2014), (Pereltsvaig, Lyutikova 2014) and (Lyutikova, Pereltsvaig, 2015) for an application of this analysis to DOM and differential possessor marking in the Mishar dialect of Tatar).

With this in mind, we are turning now to Tatar data. In the next section, we will determine case variation contexts and identify the factors licensing the variation.

4. Differential case marking in Tatar

In Tatar, at least the following contexts of morphologically observable case variation can be identified:

- Postpositional phrase complement
- Direct object

- Possessor within the noun phrase
- Subject of nominalizations
- Subject of relative clauses
- Subject of an embedded finite clause with *дип*.

The easiest case alternation to observe is with (denominal) postpositions that combine with genitive-marked personal pronouns (12a) and morphologically-unmarked nouns (12b) (in this section we will dub the morphologically unmarked form "nominative", although we do not commit ourselves to any *a priori* analysis and use this term as a purely descriptive label).

(12) a. Бел-еп тор: кайда гына бул-са-м
know-CONV AUX where EMPH be-COND-1SG
да, җан-ым гел **синең** **ян-ың-да**.
EMPH soul-1SG always you.GEN near-2SG-LOC
'Remember: wherever I am, my soul is always near you'. (http://corpus2.tatfolk.ru/)

b. Алар иртәгә **Гөлнур-ның** **әни-се**
they tomorrow Gulnur-GEN mother-3
ян-ы-на кайт-ачак-лар.
near-3-DAT return-FUT-PL
'Tomorrow they will return to Gulnur's mother'. (http://corpus2.tatfolk.ru/)

As (12a-b) demonstrate, both genitive and nominative case assignment by a preposition is accompanied by agreement. No positional or interpretational effects of differential case marking can be observed.

Differential direct object marking in Tatar has received much attention in the previous literature. A comprehensive grammatical description of Tatar (Zakiev, 1995: 119) considers Tatar DOM to be motivated by definiteness of the object NP: definite noun phrases receive accusative, whereas indefinite noun phrases are nominative. However, this interpretational effect is only observable with a subset of noun phrases, namely, those that allow for both accusative and nominative forms. In fact, there are several types of noun phrases that obligatorily receive accusative, regardless of their definiteness or even referentiality. Thus, pronominal objects (except for *нәрсә* 'what') and ezafe-3 direct objects necessarily bear the accusative case marker, even if they are indefinite and non-referential.

(13) a. Алар Муса җәлил белән Тукай-дан ары
they Musa Jalil with Tukay-ABL except
бер-кем-не дә бел-ми.
one-who-ACC EMPH know-NEG.PRS
'They don't know anyone except for Musa Jalil and Tukay'.
(http://web-corpora.net/TatarCorpus/search/)

b. Алар монда **кем-не-дер** эзли-ләр
they here who-ACC-INDEF search.PRS-PL
бул-са кирәк.
be-COND MODAL
'Apparently, they are looking for someone here'.
(http://web-corpora.net/TatarCorpus/search/)

c. Марат Алсу-ның бер **фотография-се-н** дә
Marat Alsu-GEN one picture-3-ACC EMPH
күр-мә-де.
see-NEG-PST
'Marat didn't see any picture of Alsu'.

(Lyutikova, 2014) and (Lyutikova, Pereltsvaig, 2015) suggest that the relevant factor here is the syntactic category of the noun phrase. Full-fledged noun phrases (i.e. DPs) receive the accusative case, whereas structurally deficient Small Nominals (i.e. NPs) remain unmarked. As pronouns and ezafe-3 constructions are necessarily DPs, they are obligatorily accusative. The definiteness effect of the accusative marking in other types of noun phrases results from the mobility of DPs, which can raise at LF and acquire wide scope.

It is interesting that unlike in Sakha, in Tatar accusative direct objects do not have to move out of their VP: accusative direct objects are perfectly grammatical in (14a-b), in their base position. Nominative direct objects must be strictly adjacent to the verb in both languages.

(14) a. Байрас кат-кат **хат-ны** укы-ды,
Bayras again-again letter-ACC read-PST
ни-дер аңла-рга тырыш-ты.
what-INDEF understand-INF try-PST
'Bayras read the letter again and again, trying to understand anything'. (http://web-corpora.net/TatarCorpus/search/)

b. Абы-ем-а **машина-ны** бүләк ит-әргә тели-м.
brother-1SG-DAT car-ACC present-INF want-1SG
'I want to give the car as a present to my brother'.

The distribution of genitive and nominative possessors (ezafe-3 and ezafe-2 constructions in (Zakiev, 1995)) resembles the distribution of accusative and nominative direct objects in many respects. Thus, pronouns and ezafe-3 noun phrases can be embedded only as genitive possessors (15a-b); genitive and nominative possessors clearly contrast in referentiality (15c-d).

(15) a. минем || *мин йорт-ым
I.GEN || I house-1SG
'my house'

b. Гөлнур-ның әни-се-нең || *Гөлнур-ның әни-се йорт-ы
Gulnur-GEN mom-3-GEN || Gulnur-GEN mom-3 house-3
'Gulnur's mother's house'

c. бала-лар-ның китаб-ы
child-PL-GEN book-3
'the children's book'

d. бала-лар китаб-ы
child-PL book-3
'children's book'

At the same time, the distribution of genitive and nominative possessors is more intricate. First of all, they occupy clearly different linear positions within the noun phrase: the genitive possessor is the leftmost constituent of the noun phrase, whereas the nominative possessor is the rightmost phrasal modifier of the head noun. The two ezafe constructions can co-occur within the same noun phrase, as in (16). Therefore, the two possessors not only end up occupying different surface positions, but are also necessarily base-generated in different structural positions.

(16) a. әдип-ләр-нең бала-лар күңел-е
writer-PL-GEN child-PL mind-3
'the writers' childish mind' (http://web-corpora.net/TatarCorpus/search/)

b. без-нең космос кораб-лар-ыбыз
we-PL space ship-PL-1PL
'our spaceships' (http://web-corpora.net/TatarCorpus/search/)

The next thing to observe is that noun phrases embedded under ezafe-3 and ezafe-2 constructions differ in their argumenthood: ezafe-3

hosts arguments (possessors, agents, owners, themes) and ezafe-2 hosts adjuncts. Interestingly, the nominative noun phrase in ezafe-2 construction can be even a proper name, if used in the attributive function, cf. (17a-b) and (17c):

(17) a. Татарстан республика-сы
Tatarstan republic-3
'the Republic of Tatarstan'
b. Марат урам-ы
Marat street-3
'Marat street' (i.e. the street named after Marat)
c. Марат-ның урам-ы
Marat-GEN street-3
'Marat's street' (i.e. the street where Marat lives)

Genitive possessors always trigger agreement on the head noun; however, the status of the possessive affix in ezafe-2 construction is controversial. On the one hand, it is identical with the 3rd person possessive marker of ezafe-3 construction. On the other hand, as 1st and 2nd person noun phrases are necessarily personal pronouns, they are disallowed in ezafe-2 construction, so it remains unclear whether the possessive affix in ezafe-2 reflects agreement with the 3rd person nominative possessor or is a default form. Note also that ezafe-2 construction can lose the possessive affix when embedded under ezafe-3 (as in example (16b)) or under the attributivizer *–лы* (as in *татар телле балалар* 'Tatar-speaking children').

Tatar nominalized clauses exhibit case marking variation, too: the embedded subject can receive genitive or nominative, as in (18).

(18) a. Тыңла-р иде-м кош-лар-**ның** || кош-лар
listen-FUT AUX-1SG bird-PL-GEN || bird-PL
җырла-в-ы-н.
sing-NML-3-ACC
'I would listen to the birds' singing'. (Yandex search)
b. Алаша-быз-**ның** || алаша-быз умыр-ып-умыр-ып
gelding-1PL-GEN || gelding-1PL gulp-CONV-gulp-CONV
башак аша-ган-ы-н ярат-ып, сөен-еп
mash eat-PFCT-3-ACC love-CONV rejoice-CONV
кара-п тора-м.

look-CONV AUX-1SG
'With love and joy I'm looking at our gelding gulping mash'.
(http://web-corpora.net/TatarCorpus/search/)

Although the construction may look identical to ezafe-3 vs. ezafe-2, it differs in many significant respects. First, noun phrases of any structural type (including personal pronouns and ezafe-3 DPs) exhibit the case alternation in this context. Secondly, genitive and nominative subjects of nominalized clauses, unlike possessors, do not seem to appear in distinct structural positions. Thirdly, both genitive and nominative subjects are clearly argumental.

Interestingly, nominative subjects of nominalizations, like nominative possessors in ezafe-2 constructions, allow the possessive marker on the head to drop in certain contexts (19).

(19) Алар **кит-үгә**, авыл җиңел сула-п куй-ды.
they leave-NML-DAT village easy breathe-CONV AUX-PST
'When they had left, the village people felt relieved'.
(http://web-corpora.net/TatarCorpus/search/)

Another type of embedded clauses where subjects can be nominative or genitive is participial relative clauses. Nominative subjects never trigger agreement on either the participle or head noun (20a), whereas genitive subjects require possessive agreement marker on the head noun (20b).

(20) a. Марат Казан-нан ал-ып кайт-кан
Marat Kazan-ABL take-CONV return-PFCT
китап бик кызык.
book very interesting
'The book that Marat brought from Kazan is very interesting'.
b. Марат-**ның** Казан-нан ал-ып кайт-кан
Marat-GEN Kazan-ABL take-CONV return-PFCT
китаб-**ы** бик кызык.
book-3 very interesting
'The book that Marat brought from Kazan is very interesting'.

It seems that the genitive case marking is only available if the subject occupies the leftmost position in the relative clause; thus, inter-

changing the NPs *Маратның* ‘Marat’s’ and *Казаннан* ‘from Kazan’ in (20b) results in ungrammaticality.

Finally, subjects of embedded finite clauses headed by *дип* show case marking variation between nominative and accusative. Both forms control agreement on the embedded predicate (21a). The accusative subject seems to be structurally higher than the nominative one – as in the previous case, the subject NP has to be the leftmost constituent of the embedded clause in order to receive the accusative case marking. But, unlike in Sakha, only transitive matrix verbs in Tatar license accusative case marking of the embedded subject, cf. (21a-b), (22a-b).

(21) a. Син || сине кил-ер-сең дип көтә-м.
you || you.ACC come-FUT-2SG that wait.PRS-1SG
‘I’m waiting for you to come’.

b. Син || *сине кил-ер-сең дип көт-ел-ә.
you || you.ACC come-FUT-2SG that wait-PASS-PRS
‘It is expected that you would come’.

(22) a. Алсу Марат || ʔМарат-**ны** кунак-ка кил-ер
Alsu Marat || Marat-ACC guest-DAT come-FUT
дип өй-не җыештыр-ды.
that home-ACC clean-PST
‘Alsu cleaned the house because Marat should come to visit’.

b. Алсу Марат || *Марат-**ны** кунак-ка кил-ер
Alsu Marat || Marat-ACC guest-DAT come-FUT
дип бизән-де.
that make_up-PST
‘Alsu did makeup because Marat should come to visit’.

Let us take stock of what we have learned so far. In all the contexts of morphologically observable case variation, one of the varying case forms is always nominative. Genitive noun phrases always trigger possessive agreement, nominative subjects always trigger predicative agreement in finite clauses and possessive agreement in some nominalized clauses, nominative possessors condition the presence of the possessive marker on the head noun (its status of an agreement marker being unclear); accusative subjects control predicative agreement in the finite embedded clause; nominative and accusative direct objects show no connection with whatever agreement.

The clear positional distinction is found in half of the contexts – namely, with genitive/nominative possessors, genitive/nominative subjects of relative clauses and accusative/nominative subjects of embedded finite clauses with *дип*. Accusative/nominative direct objects are not positionally distributed in the strict sense of the term, but they differ as to the positions available: the nominative direct object has to be adjacent to its verb whereas the accusative direct object may move or stay in its base position.

As for functional heads assigning structural cases, we can clearly identify finite T as a nominative case assigner, possessive D as a genitive case assigner, transitive *v* as an accusative case assigner and P as a genitive/nominative case assigner. It remains unclear whether nominative subjects of non-finite clauses, nominative direct objects and nominative possessors are case-depending on some functional head.

Finally, the relevant factors underlying case marking variation differ to a large extent – from purely formal pronoun/noun distinction to intricate argumental/attributive semantic type variability, definiteness, or referentiality. Sometimes, however, it seems almost impossible to reveal any semantic impact of differential case marking.

The relevant data are summarized in Table 2.

Table 2

Differential case marking in Tatar

Context type	Variation	Agreement	Positional variation	Case-assigning F	Licensing factors
PP complement	GEN/NOM	+/+	–	P/P	pronoun/noun
Direct object	ACC/NOM	–/–	±	*v*/?	DP/SN definite/ indefinite
Possessor	GEN/NOM	+/?	+	D/?	DP/SN referential/non-referential argument/ adjunct

Subject of nominalization	GEN/NOM	+/±	–	D/?	?
Subject of relative clause	GEN/NOM	+/–	+	D/?	?
Subject of *∂un*-clause	ACC/NOM	+/+	+	*v*/T	?

5. Discussion

In the previous section we have identified syntactic constructions where differential case marking is available in Tatar. We have observed that Tatar differs from Sakha in several significant respects. First, it allows for both genitive and nominative possessors within the noun phrase; secondly, its accusative and nominative direct objects are not fully distributed positionally; thirdly, embedded subjects can bear nominative morphology in non-finite clauses, and the agreement with such subjects is not obligatory; fourthly, the accusative embedded subject is causally dependent on the transitivity of the matrix clause.

Neither Levin and Preminger's analysis, nor Baker and Vinokurova's analysis of case assignment in Sakha can be extended to Tatar data. The obvious problem is the accusative case assignment, which is determined by the presence of the transitive functional head, and not by the presence of another NP in the clausal domain. Besides, the accusative direct object does not have to leave its VP, and therefore both accusative and nominative objects belong to the same domain of case assignment.

Another problem is the availability of both genitive and nominative possessors. Since the two different cases are assigned in the same domain, the "NP's unmarked case" analysis of the genitive cannot be maintained. Note that neither genitive nor nominative in the nominal domain can be characterized as "dependent" case, because neither ezafe-2 is causally dependent on ezafe-3, nor vice versa. The Chomsky-style analysis of the genitive proposed by Baker and Vinokurova would fit the data nicely, but the nominative possessor then requires its own licenser and case assigner.

The most controversial is certainly the nominative case distribution. It cannot be "the clause's unmarked case", because it can be found in

different domains – verbal, nominal, prepositional. It cannot be a default case, because it is unavailable for pronouns in prepositional phrases and for DPs in the direct object position or the possessor position. It cannot be always assigned by an agreeing functional head, because it sometimes must trigger agreement (as a finite subject), sometimes can trigger agreement (as a subject of a nominalization) and sometimes cannot trigger agreement (as a relative clause subject).

We therefore think that Tatar data requires a hybrid approach to the case assignment, but not in the sense of mixing Chomsky-style and Marantz-style modalities in one model. Instead, we propose a purely syntactic account that distinguishes between different "nominatives".

Following (Lyutikova, 2014), (Pereltsvaig, Lyutikova, 2014), (Lyutikova, Pereltsvaig, 2015), we claim that in Tatar, noun phrases differ as to their structural complexity, so that fully projected DPs like pronouns or ezafe-3 nominals coexist with structurally deficient Small Nominals (SNs). We believe that only DPs are subjects to the Case filter and have to receive abstract case in order to be licensed, and SNs can go caseless. So we suppose that nominative direct objects and nominative possessors are in fact caseless SNs.

Nominative subjects, on the other hand, are clearly case-marked, since the full-fledged DPs are readily available in this position. We believe that finite and non-finite clauses in Tatar differ as to the morphological spell-out of T rather than its featural specification. So we hypothesize that T always assigns nominative to the subject under AGREE, but the results of the predicative agreement are only visible in finite clauses.

Finally, the nominative form of the prepositional phrase complement is clearly neither a caseless form (as it is available for ezafe-3 nominals) nor a subject case. Since the distribution of genitive and nominative forms is purely morphological, we propose that they are morphological exponents of an abstract "postpositional" case, which could then be characterized as "morphologically non-independent" in terms of (Zaliznyak, 1973: 74).

Other case forms correspond directly to case assigning functional heads. So the adnominal genitive is assigned by the agreeing functional head D, and the accusative is the case of the transitive *v*. The resulting specification of Tatar case forms is given in Table 3.

Table 3

Tatar structural cases in Chomsky-style case assignment model

Case form/Context	Abstract case	Case-assigning functional head	Agreement
NOM/finite clause subject	NOM	T finite	+ (predicative)
NOM/non-finite clause subject	NOM	T non-finite	– (morphologically invisible)
NOM/noun phrase possessor	–	–	– (ezafe-2 possessive marker)
NOM/direct object	–	–	–
NOM/PP complement	POSTP	P	+ (possessive)
GEN/PP complement	POSTP	P	+ (possessive)
GEN/elsewhere	GEN	D	+ (possessive)
ACC	ACC	*v* transitive	– (morphologically invisible)

The proposed hybrid account combines Chomsky-style case assignment model with two additional ideas, namely, that some nominals can be caseless and that the spell-out of the abstract case can be morphologically conditioned. It seems that this account is superior to both the purely configurational model of Levin and Preminger and the hybrid model of Baker and Vinokurova, as it allows us to deal with Tatar case marking variation phenomena in a uniform and comprehensive manner.

Acknowledgements

The authors gratefully acknowledge the financial support of Russian Scientific Foundation to Ekaterina Lyutikova (project # 14-18-03270 ‘Word order typology, communicative-syntactic interface and information structure in world’s languages’) and the financial support of Russian Scientific Foundation for Humanities to Dilya Ibatullina (project # 14-04-00580 ‘Interaction of grammatical mechanisms in world’s languages’). Our warmest thanks to Asya Pereltsvaig for her invaluable

help and inspiring discussion, and to the conference organizers, especially Dzhavdet Suleymanov and Olga Nevzorova, who encouraged us to present our results.

REFERENCES

Aissen, J. (2003). Differential Object Marking: Iconicity vs. Economy. *Natural Language and Linguistic Theory* 21:435–483.

Baker, M. & Vinokurova, N. (2010). Two modalities of Case assignment: Case in Sakha. *Natural Language and Linguistic Theory* 28:593–642.

Baker, M. (2009). Is head movement still needed for noun incorporation? The case of Mapudungun. *Lingua* 119(2):148–165.

Bittner, M. & Hale, K. (1996). The structural determination of Case and agreement. *Linguistic Inquiry* 27:1–68.

Bobaljik, J. (2008). Where's phi? Agreement as a post-syntactic operation. In D. Harbour et al. (eds.). *Phi Theory*. (pp. 295–328). Oxford: OUP.

Bossong, G. (1985). *Differentielle Objektmarkierung in den Neuiranischen Sprachen*. Tübingen: Gunter Narr Verlag.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

Chomsky, N. (1986). *Knowledge of language*. New York: Praeger.

Chomsky, N. (2000). Minimalist inquiries: the framework. In R. Martin, D. Michaels and J. Uriagereka (eds.). *Step by Step. Essays on Minimalist Syntax in Honor of Howard Lasnik.* (pp. 89–155). Cambridge, Mass.: MIT Press.

Chomsky, N. (2001). Derivation by phase. In M. Kenstowicz (ed.). *Ken Hale: A life in language*. (pp. 1–52). Cambridge, Mass.: MIT Press.

Danon, G. (2006). Caseless nominals and the projection of DP. *Natural Language and Linguistic Theory* 24:977–1008.

Kornfilt, J. (1997). *Turkish grammar*. London: Routledge.

Kornfilt, J. (2013). *Nominative as no case at all: An argument from raising-to-accusative in Sakha*. Paper presented at WAFL9 workshop, August 23–25, 2013, Cornell University, Ithaca, NY.

Kornfilt, J., & Preminger, O. (2014). *Nominative as no case at all: an argument from raising-to-accusative in Sakha*. Ms., Syracuse University. [URL: http://omer.lingsite.org/files/Kornfilt-and-Preminger---NOM-as-no-case-at-all---Sakha.pdf]

Kurilovich, E. (1962). The problem of case classification [Problema klassifikatsii padezhey]. In Kurilovich E. *Essays in Linguistics* [Ocherki po lingvistike] (pp. 175–203). Moscow: Nauka.

Levin, T. & Preminger, O. (2015). Case in Sakha: Are two modalities really necessary? *Natural Language and Linguistic Theory* 33(1):231–250.

Lyutikova, E. A. & Pereltsvaig, A. M. (2015). Noun phrase structure in articleless languages: universality and variation [Struktura imennoy gruppy v bezartiklevyx yazykax: universal'nost' i variativnost']. *Voprosy yazykoznaniya* 3:52–69.

Lyutikova, E. A. (2014). Case and noun phrase structure: differential object marking in Mishar dialect of Tatar [Padezh i struktura imennoy gruppy: variativnoye markirovaniye ob'ekta v misharskom dialekte tatarskogo yazyka]. *Vestnik MGGU im. M. A. Sholokhova, Ser. Filologicheskiye nauki* 4:50–70.

Marantz, A. (1991). Case and licensing. In G. Westphal, B. Ao and H. Chae (eds.). *Proceedings of the 8th Eastern States Conference on Linguistics (ESCOL 8)*. (pp. 234–253). Ithaca, NY: CLS Publications.

Massam, D. (2001). Pseudo noun incorporation in Niuean. *Natural Language and Linguistic Theory* 19:153–197.

McFadden, T. (2004). *The position of morphological case in the derivation*. Doctoral dissertation, Philadelphia, PA: University of Pennsylvania.

Moravcsik, E. (1974). Object-verb agreement. In *Working papers on language universals* 15:25–140.

Moravcsik, E. (1978). Agreement. In J. Greenberg (ed.). *Universals of human language IV: syntax.* (pp. 331–374). Stanford, CA: Stanford University Press.

Pereltsvaig, A. & Lyutikova, E. (2014). Possessives within and beyond NP: Two ezafe-constructions in Tatar. In A. Bondaruk, G. Dalmi and A. Grosu (eds.). *Advances in the syntax of DPs: Structure, agreement, and case*. (pp. 193–219). Amsterdam: Benjamins.

Pereltsvaig, A. (2006). Small nominals. *Natural Language and Linguistic Theory* 24:433–500.

Preminger, O. (2011). *Agreement as a fallible operation*. Doctoral dissertation, Cambridge, MA: MIT.

Zakiev, M. Z. (1995). *Tatar Grammar, Vol. 3: Syntax* [Tatarskaya grammatika, Tom III: Sintaksis]. Kazan: Akademiya Nauk Tatarstana.

Zaliznyak, A. A. (1973). On the term Case in linguistic descriptions [O ponimanii termina padezh v lingvisticheskix opisaniyax]. In *Problems of grammatical modelling* [Problemy grammaticheskogo modelirovaniya] (pp. 53–88). Moscow: Nauka.

SECTION 6 | **SPEECH TECHNOLOGIES**

ABOUT METHODS ESTIMATING THE PARAMETERS OF SIGNALS[1]

Dmitry Alyunov

«Chuvash State University n.a. I.N. Ulianov», Cheboksary, Russia
(428015, Cheboksary, Moscow Prospect,15)
e-mail: aldmitry89@gmail.com

There are ways of estimating the power spectral density of the signals, their use for the evaluation of speech parameters. The drawbacks of classical methods of spectral estimation – their dependence on the length of the analyzed signal, the effects of the spreading of the spectrum, a property exchange frequency resolution on the smoothness of assessment, the use of windows for smoothing the spectrum, especially the windows, the dependence of the quality of estimation of the percentage of overlapping windows. The advantages and disadvantages of parametric methods (Berg, covariance method, the modified covariance method), their features – trickle-down effect, masking weak signal stronger, their advantages – the ability to allocate the necessary components of the signal at shorter intervals, compared to the classical methods. The approach segment definitions of words in continuous speech, which allows a high degree of accuracy to identify words with a high level of noise and background.

Распознавание речи с каждым годом находит все большее применение в нашей жизни(цифровая передача и хранение данных, синтез речи, идентификация диктора, устранение дефектов речи, улучшение параметров речевого сигнала), разрабатываются новые алгоритмы. Разработаны модели, описывающие речь ее информационным содержанием, модели представляющие речь сигналом и

[1] Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 14-07-00143

т.п. Обработка сигнала представляет собой формирование описания на основе некоторой модели с последующим преобразованием полученного представления. Заключительным шагом является выделение необходимых параметров сигнала, их анализ и работа с ними.

Исторически распространенными классическими методами спектрального оценивания являются периодограммный и коррелограммный методы оценивания спектральной плотности мощности (СПМ) сигнала.

Спектральная плотность мощности (СПМ) стационарного случайного процесса есть дискретно-временное преобразование Фурье (ДВПФ) автокорреляционной последовательности (r_{xx}).

$$P_{xx}(f) = T \sum_{m=-\infty}^{\infty} r_{xx}[m] e^{-j2\pi fmT} \quad (1)$$

Если допустить, что процесс является эргодическим, то:

$$P_{xx}(f) = \lim_{M\to\infty} \left\langle \frac{1}{(2M+1)T} \left| T \sum_{n=-M}^{M} x[n] e^{-j2\pi fnT} \right|^2 \right\rangle \quad (2)$$

Если определять спектральную плотность мощности исходя из автокорреляционной последовательности, получится коррелограммный метод, поскольку случайный процесс непосредственно не используется для оценки СПМ. В том случае, если использовать саму числовую последовательность для оценки СПМ, получится периодограммный метод.

Оценка СПМ, получаемая на основе коррелограммного метода принимает форму:

$$P_{xx}(f) = T \sum_{m=-L}^{L} w[m] \hat{r}_{xx}[m] e^{-j2\pi fmT} \quad (3)$$

Среднее значение этой оценки будет сверткой истинного спектра и спектра окна W(f) :

$$\langle \hat{P}(f) \rangle = T \sum_{m=-L}^{L} w[m] r_{xx}[m] e^{-j2\pi fmT} = P_{xx}(f) * W(f) \quad (4)$$

Правильный выбор окна позволит уменьшить растекание спектра и его смещение.

При наличии конечного множества данных x(n), $0 \le n \le N-1$ и единственной реализации, это соотношение преобразуется в СПМ выборки или периодограмму:

$$\widetilde{\widetilde{P_{xx}}}(f) = \frac{1}{NT} \left| T \sum_{n=0}^{N-1} x[n] \; e^{-j2\pi fnT} \right|^2 = \frac{T}{N} \left| \sum_{n=0}^{N-1} x[n] e^{-j2\pi fnT} \right|^2. \quad (5)$$

На рисунке (1) представлены оценки СПМ периодограммным методом Уэлча без усреднения и с усреднением по 18 сегментам соответственно. Усредненная оценка имеет гораздо меньшую дисперсию, однако пришлось пожертвовать разрешением спектральных компонент. На графике слева изображена оценка СПМ суммы двух комплексных экспонент с относительными частотами 0.35 и 0.36 периодограммой Уэлча, 1 сегмент – длина окна 1000 отсчетов, SNR=6. Справа – периодограмма Уэлча, 18 сегментов – длина окна 100 отсчетов, сдвиг сегмента 50 отсчетов, SNR=6.

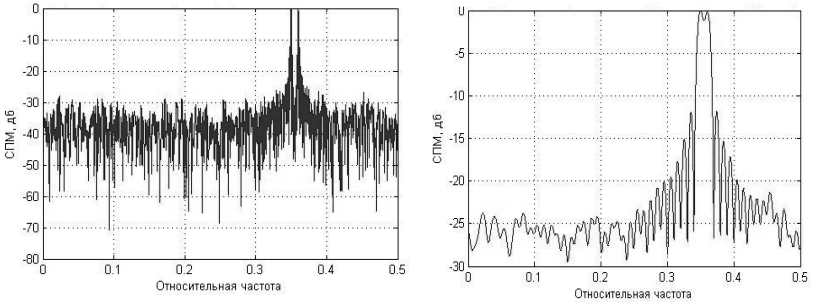


Рис.1. Влияние усреднения на качество периодограммного метода оценивания СПМ

В данном примере использовалось прямоугольное окно. Ширина его главного лепестка уже чем, например, у окна Хэмминга, поэтому и разрешение спектральных компонент эффективнее. Применение окна уменьшает эффекты просачивания и маскировки вследствие того, что уширение главного лепестка спектра окна происходит за счет уменьшения уровней боковых лепестков. Качество разрешения зависит от количества взятых отсчетов. Аналогичная ситуация и в коррелограммных оценках СПМ близких гармоник. Но имеются некоторые особенности.

Многие случайные процессы дискретного времени описываются следующей моделью:

$$x[n]=-\sum_{k=1}^{p} a[k]x[n-k]+\sum_{k=0}^{q} b[k]u[n-k]=\sum_{k=0}^{\infty} h[k]u[n-k] \qquad (6)$$

Это можно представить в виде выхода фильтра, где x[n] – входная последовательность каузального фильтра, u[n] – входная возбуждающая последовательность(белый шум с нулевым средним и

дисперсией ρ_w), h[n] –импульсная характеристика фильтра, a[k] – коэффициент авторегрессии, b[k] – коэффициент скользящего среднего.

Процесс на выходе фильтра (8) соответствует модели авторегрессии – скользящего среднего (АРСС), где параметры a[k] характеризуют авторегрессионную часть этой модели порядка p, а параметры b[k] – ее часть, соответствующую скользящему среднему порядка q.

Спектральная плотность мощности для АРСС процесса имеет вид

$$P_{\text{АРСС}}(f) = T\rho_w |B(f)A(f)|^2 \quad (7)$$

$$\text{A(f)}=1+\sum_{k=1}^{p} a[k]\exp(-j2\pi kTf), \quad (8)$$

$$\text{B(f)}=1+\sum_{k=1}^{q} b[k]\exp(-j2\pi kTf), \quad (9)$$

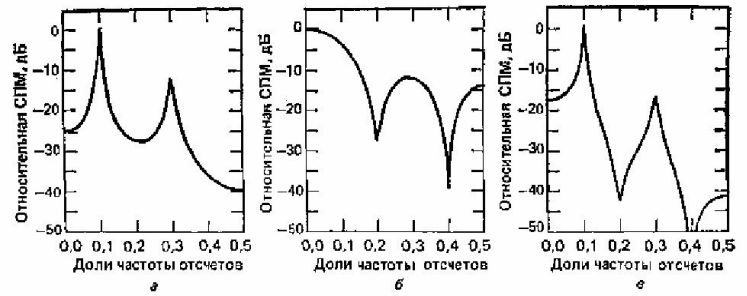


Рис.2. Примеры оценивания СПМ сигнала с использованием параметрических методов

Первый из параметрических методов – метод Берга, гораздо точнее оценивал результаты. Но он был не без недостатков: расщепление спектральных, эффект смещения спектральных, зависящий от начальной фазы гармоник. В ковариационных методах эти недостатки были устранены (рис. 3 слева – сигнал смеси 2-х синусоид с относительными частотами 0.3 и 0.34; SNR=30). Данные параметрические методы получили широкое распространение в ЦОС, поскольку они позволяют получать высокое разрешение и острые спектральные пики спектральных компонент. При низком порядке модели получаются более сглаженные спектральные оценки, при излишне высоком – увеличивается разрешение, но в спектре появляются ложные пики.

Авторегрессионные методы имеют гораздо лучшее разрешение спектральных компонент по сравнению с классическими методами спектрального оценивания. На рисунке (3) справа очевидно превосходство параметрических алгоритмов ЦСА. Красной линией изображена периодограмма а синей линией – модифицированный ковариационный метод. Длина последовательности выбрана в 30 отсчетов. Спектр сигнала состоит из 2 спектральных пиков на частотах 0.3 и 0.35.

При использовании классических алгоритмов спектрального оценивания следует учитывать произведение «устойчивость*дли тельность*ширина полосы»; имеет место свойство обмена частотного разрешения на гладкость оценки; характерны эффекты маскирования и растекания спектра.

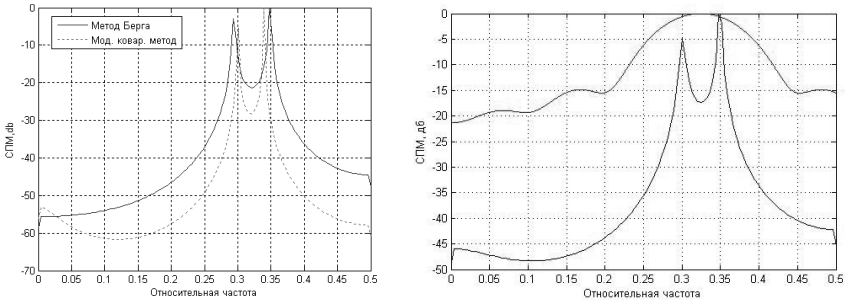


Рис.3. Сравнение работы модифицированного ковариационного метода и метода Берга – слева. Сравнение качества оценивания двух спектральных компонент – справа

Увеличение порядка АР - модели сопровождается улучшением частотного разрешения, однако при избыточном порядке модели возникают ложные спектральные пики; для всех анализируемых методов характерно следующее свойство: при увеличении числа анализируемых отсчетов сигнала или порядка модели частотное разрешение повышается, однако дисперсия оценки СПМ увеличивается.

Интересным выглядит изучение информационной энтропии спектра сигнала. Как известно, энтропия шумового сигнала и речевого сигнала отличается, и, что можно отнести к преимуществам данного метода, энтропия мало чувствительна к амплитуде сигнала. Принцип работы изображен на Рисунке (4).

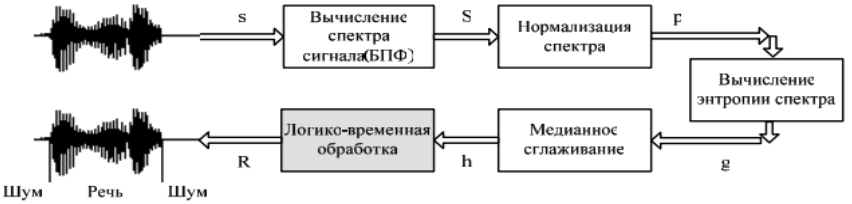


Рис. 4. Анализ речевого сигнала с использованием энтропии спектра. Алгоритм

Как происходит обработки сигнала: сигнал дискретизируется, затем делится на сегменты по 256 цифровых отсчетов, перекрытие сегментов сделаем немногим более 25% для устранения краевых сегментов. Мгновенные спектры мощности сигнала рассчитываем по следующей формуле:

$$S_i(k) = \left| \sum_{m=0}^{M-1} s_i(m) e^{-j\frac{\pi}{M}mk} \right|^2, \quad 0 \le k \le M, \qquad (10)$$

M – размер сегмента сигнала.

Дальше нормализуем спектр по всем частотным компонентам:

$$p_i = \frac{S(f_i)}{\sum_{k=1}^{N} S(f_k)} \qquad (11)$$

Таким образом мы получили плотность вероятности спектра. Мы ее ограничиваем: верхним и нижним пределом. Если есть равномерное распределение частотных компонент – это белый шум, а также мы исключаем шумы в узкой частотной области.

$$p_i = \begin{cases} 0, p_i > \delta_1 \\ 0, p_i < \delta_2 \\ p_i, иначе \end{cases}, \qquad (12)$$

δ_1 и δ_2 – верхний и нижний пределы.

Из экспериментальных расчетов можно оценить порядок пределов как 0,01 и 0,3 при p от 0 до 1. Кроме того, в дополнение

мы можем использовать и другие способы выделения сигнала из шума: методы спектрального вычитания и адаптивный фильтр Кальмана.

Мы рассматриваем энтропию как меру беспорядка в распределении, рассчитывая по следующей формуле:

$$H = -\sum_{k=1}^{N} p_k \log p_k \,. \tag{13}$$

Полученную функцию необходимо сгладить, используя медианное сглаживание. Данный тип сглаживания является наиболее устойчивым по отношению к случайным выбросам. Для этого берется какой либо интервал [t-q,t+q] и вычисляется скользящая медиана в точке t [6]. Медиана ряда интервале определяется как центральный член последовательности значений ряда, входящих в этот временной интервал, упорядоченной по возрастанию. Как показывают эксперименты наиболее точное вычисление медианы происходит при окне величиной в 5 сегментов. В том случае, когда мы вычисляем медиану в точках, близких к краю интервала и меньших чем размер окна, приходится уменьшать окно.

Затем мы вычисляем порог для определения границ речевого сигнала.

$$r = \left(\frac{\max(h) - \min(h)}{2} + \min(h) \right) * \mu \tag{14}$$

, μ - коэффициент зашумленности.

Он подбирается экспериментально, зависит от параметров шума. Данный коэффициент может принимать значения от 0,8 до 1,1 в зависимости от уровня шума. На основе вычисленного значения r выбираются акустические сегменты речевого сигнала.

Заключительным этапом является логическая временная обработка полученной энтропии спектра, используя допустимые на практике длительности речевых и неречевых сигналов, вычисленных ранее. Это необходимо, поскольку зачастую различные звуковые эффекты(кашель и прочее) принимают за речь, а некоторые участки речи за межречевой интервал. Используя адаптивный порог можно определить сегменты речи на основе вычисления максимальной длительности межречевого участка S, и минимальной длительности участка речи (Рисунок 2).

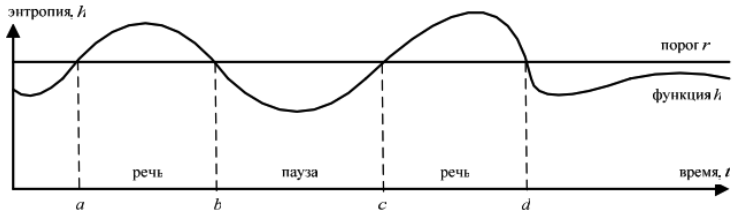


Рис. 5. Энтропия и порог обнаружения

Поскольку человек чисто физически не может произносить речевые фрагменты короче определенного значения, и так как всегда присутствуют паузы, можно экспериментально определить значения параметров R и S. Анализирем полученные результаты следующим образом:

$$речь = \begin{cases} ad, (ab \geq R) \wedge (cd \geq R) \wedge (bc \leq S) \\ ab, (ab > cd) \wedge (ab \geq R) \\ cd, (ab < cd) \wedge (cd > R) \\ \varnothing, иначе \end{cases} . \qquad (20)$$

Если участок сигнала без речи содержит не более S сегментов, а участки, содержащие речь составляют не менее R сегментов, то образуется сплошной речевой участок.

Экспериментальные расчеты показывают, что наименьший процент ошибок получается в случае узкополосного или белого шума – порядка 1,5%. Наихудший результат – 16% ошибок в случае розового шума, наиболее приближенного к реальной речи. К минусам данного метода можно отнести тот факт, что фоновая речь, пусть даже слабая, может быть принята за полезный сигнал. В целом данный метод довольно неплохо определяет речь в сигналах с высоким уровнем шумов и нестационарности.

СПИСОК ЛИТЕРАТУРЫ.

1. Гоноровский И.С. Радиотехнические цепи и сигналы: учебник для вузов – 4-е изд., перераб. И доп. – М.: Радио и связь, 1986;

2. Бияков, О.А. Медианное сглаживание временных рядов / О.А. Бияков // Вестник КузГТУ. 1999. №3. С. 55–56.

3. Кривошеев В.И. Современные методы циф-ровой обработки сигна-

лов (цифровой спектральный анализ): Учебно-методические материалы. Н. Новгород: ННГУ, 2006. 117 с. http://www.unn.ru/pages/e-library/aids/2006/7.pdf.

4. Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов // М.:Радио и связь, 1981 – 496 с.

5. Цымбал, В.П. Теория информации и кодирование / В.П. Цымбал // Киев.: Высшая Школа, 1977, 288 с.

6. Waheed, K. A robust algorithm for detecting speech segments using an entropy contrast / K. Waheed, K. Weaver, F. Salam // Proceedings of MWSCAS'2002, Oklahoma, USA, 2002.

7. Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов // М.:Радио и связь, 1981 – 496 с.

8. Кэй С.М., Марпл С.Л. Современные методы спектрального анализа: Обзор // ТИИЭР. 1981. Т. 69. № 11. С. 5–50.

Рецензенты: Артемьев Иосиф Тимофеевич, д.ф.-м.н., профессор, зав. кафедрой Кафедра математического и аппаратного обеспечения информационных систем (МиАОИС) ФГБОУ ВПО «Чувашский государственный университет им. И.Н. Ульянова», город Чебоксары, Охоткин Григорий Петрович, Декан факультета радиоэлектроники и автоматики (ФРЭА), д.т.н., профессор ФГБОУ ВПО «Чувашский государственный университет им. И.Н. Ульянова», город Чебоксары.

KAZAKH WORDS RECOGNITION BASED ON DIPHONE DATABASE

Aygerim Buribayeva, Altynbek Sharipbay, Gulmira Bekmanova[1]

L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
[1]author@institute.xxx

In this article we propose method of word recognition based on diphones base and principles to create diphones base of Kazakh language. The system recognizes non individual diphones, whole words by reference, synthesized from diphones. Automatic generation of suggested words from diphones will make a step up towards ultra-large vocabularies.

Introduction

Automatic speech recognition of a natural language is one of the important areas development of artificial intelligence. Results in this direction will help to solve the problem of creating an effective means of verbal human-computer interaction. Speech input has a number of advantages, such as naturalness, efficiency, semantic accuracy of input, hands-free and user's perspective, the ability to manage and process in extreme conditions.

Investigation of the speech recognition problem for more than 50 years by specialists of several scientific fields. Methods and algorithms that are used can be divided into four major classes:

1. Discriminant analysis methods based on Bayesian Discrimination [1];
2. Hidden Markov Model [2];
3. Artificial neural networks [3];
4. Dynamic programming – Dynamic Time Warping (DTW) [4];

It should be mentioned several advantages are aiming at the development of speech recognition systems :

5. Continuous speech – the ability that allows users to speak naturally (continuously).
6. Large dictionaries – the ability to process large amounts of words of both general and special categories of technical and knowledge subject areas in order to increase the power and efficiency of voice recognition systems.
7. Independence from the speaker – the system's ability to recog-

nize words without a personal computer settings by repeating the same speech signal.

It frequently and successfully in continuous speech recognition uses hidden Markov model (HMM) [5, 6] or Artificial Neural Networks [6, 7]. In order to recognize the speech it selecst different basic units: phonemes, allophones, diphones, triphones etc. For individual words temporary dynamic algorithms (DTW) [8] are more effective.

Due to the fact that the recognition of whole words is more reliable, we chose words recognition technology, based on trained diphone database [9]. The point is, the system does not recognize the diphones separately. First of all it synthesizes standard words, after recognize whole words by the DTW algorithm. The advantage of the system is, in order to add a new word there is no need to train the system again, you may simply enter a word in the text. Automatic generation of standard words from diphones will make a step up towards ultra-large vocabularies. Speaker-independence of system could be achieved by averaging of standards.

2. Creating a Kazakh's diphones database

Diphone – sound unit, what has length from the middle of one sound to the middle of the next. Diphone model is based assumption. There are stationary portions of sounds, and they are independent from influence of neighboring sounds

Acoustic base of speech recognition system includes three types of diphones – initial, middle, and final.

Leading and trailing diphones usually represent the first and last half words of phonemes with inclusion transition sections from space bar to the phoneme and phonemes on to the space, respectively. They are defined according to the rules of the Kazakh positional sounds:

8. а, ә, е, ө, ұ, ү, ы, і found in all positions;

9. vowel sound о occurs only in the initial syllable;

10. л, р, й, ң, у (w) are not found in the beginning of words;

11. consonants б, д, ғ, г are not found at the end of words.

«О», «ө» and «е» are diphthongs, and according to the Kazakh orthoepy before sounds «о», «ө» which come in the beginning of words, there is a small inset consonant «у», and before «е» – consonant insertion «й». Accordingly, sounds «о», «ө» and «е» were removed from the

list of initial half diphones and instead of them were added «у» и «й» half diphones.

Taking into account all above rules, we have compiled a list of the initial and final diphones (Table 1):

Table 1

Initial and final diphones of Kazakh language

Initial		Final	
а0	м0	а2	ө2
ә0	н0	ә2	п2
б0	п0	е2	р2
г0	с0	ж2	с2
ғ0	т0	з2	т2
д0	у0	й2	у2
ж0	ұ0	к2	ұ2
з0	ү0	қ2	ү2
й0	ш0	л2	ш2
к0	ы0	м2	ы2
қ0	і0	н2	і2
		ң2	

For compiling matrix middle of diphones first automatically generated list of all possible combinations of sounds of Kazakh language. Were then removed from the list of combinations that are contrary to the following Kazakh positional rules:

12. sounds а, ә, о, ө, ұ, ү combined with all the consonants;
13. sounds е, ы, і not combined with a consonant у (w).
14. in Kazakh language are not encountered 2 consecutive vowels;
15. voiced and unvoiced consonants are not combined;
16. consonant у has not found after consonants;

Some combinations have been removed due to the fact that they do not meet at all [10].

In the end, we got about 500 sound combinations of Kazakh language. But for the qualitative recognition is not enough, as the Kazakh language is synharmonic language.

Let us consider diphones of the sound combinations in which one of the sounds is a vowel (table 2, appendix A). Their quantity remains unchanged, as determined labial labial / non-labial vowels and softness / hardness of the consonant.

Table 2

Diphones with consonant-vowel sound combination

	а	**ә**	**е**	**о**	**ө**	**ұ**	**ү**	**ы**	**і**
б	ба	б'ә	б'е	б ᵒо	б'ᵒө	б ᵒұ	б'ᵒ ү	бы	б'і
г		г'ә	г'е		г'ᵒө		г'ᵒ ү		г'і
ғ	ға			ғ ᵒо		ғ ᵒұ		ғы	
д	да	д'ә	д'е	д ᵒо	д'ᵒө	д ᵒұ	д'ᵒ ү	ды	д'і
ж	жа	ж'ә	ж'е	ж ᵒо	ж'ᵒө	ж ᵒұ	ж'ᵒ ү	жы	ж'і
з	за	з'ә	з'е	з ᵒо	з'ᵒө	з ᵒұ	з'ᵒ ү	зы	з'і
й	йа	й'ә	й'е		й'ᵒө	й ᵒұ	й'ᵒ ү	йы	й'і
к		к'ә	к'е		к'ᵒө		к'ᵒ ү		к'і
қ	қа			қ ᵒо		қ ᵒұ		қы	
л	ла	л'ә	л'е		л'ᵒө	л ᵒұ	л'ᵒ ү	лы	л'і
м	ма	м'ә	м'е	м ᵒ о	м'ᵒө	м ᵒұ	м'ᵒ ү	мы	м'і
н	на	н'ә	н'е	н ᵒ о	н'ᵒө	н ᵒұ	н'ᵒ ү	ны	н'і
ң	ңа	ң'ә	ң'е		ң'ᵒө	ң ᵒұ	ң'ᵒ ү	ңы	ң'і
п	па	п'ә	п'е		п'ᵒө	п ᵒұ	п'ᵒ ү	пы	п'і
р	ра	р'ә	р'е		р'ᵒө	р ᵒұ	р'ᵒ ү	ры	р'і
с	са	с'ә	с'е	с ᵒ о	с'ᵒө	с ᵒұ	с'ᵒ ү	сы	с'і
т	та	т'ә	т'е	т ᵒ о	т'ᵒө	т ᵒұ	т'ᵒ ү	ты	т'і
у	уа	у'ә		у ᵒ о	у'ᵒө	у ᵒұ	ұ'ᵒ ү		
ш	ша	ш'ә	ш'е	ш ᵒ о	ш'ᵒө	ш ᵒұ	ш'ᵒү	шы	ш'і

There is a question, «What about the diphones with a consonant sound combination, according to?» Is it possible, for example, the same diphone «ст» standards apply for the synthesis of the words «астау» and «үстөу»? Or apply to each: solid, non-labial «ст» for the «астау»

and soft lip «ст» for «үстөу»? And how many versions of «ст» may be at all?

We decided to to split Kazakh diphones consisting exclusively from consonants in the following groups on synharmonic Voices:

1. Hard non-labial/ Hard non-labial;
2. Hard non-labial /Soft non-labial;
3. Hard non-labial/ Hard labial;
4. Hard non-labial/ Soft labial;
5. Soft non-labial/ Hard non-labial;
6. Soft non-labial/ Soft non-labial;
7. Soft non-labial/ Hard labial;
8. Soft non-labial/ Soft labial;
9. Hard labial / Hard non-labial;
10. Hard labial / Soft non-labial;
11. Hard labial / Hard labial;
12. Hard labial / Soft labial;
13. Soft labial / Hard non-labial;
14. Soft labial / Soft non-labial;
15. Soft labial / Hard labial;
16. Soft labial / Soft labial.

Thus, the same diphone might have 16 versions (Appendix B).

But during the experiment revealed that Incorporation of all of these features for recognition is not required. For qualitative detection is sufficient to consider the softness / hardness of diphones. As a result, the basis for each diphone accordance with the consonant-sound combination left only 4 options.

Table 3

Diphones according with a consonant sound combination, with considering of softness and hardness of the components of the sound

	тн/тн	тн/мн	мн/тн	мн/мн
бд	бд	бд'	б'д	б'д'
бж	бж	бж'	б'ж	б'ж'
бз	бз	бз'	б'з	б'з'
гб	гб	гб'	г'б	г'б'
гд	гд	гд'	г'д	г'д'

гж	гж	гж'	г'ж	г'ж'
	тн/тн	тн/мн	мн/тн	мн/мн
гз	гз	гз'	г'з	г'з'
ғб	ғб	ғб'	ғ'б	ғ'б'
		…		
шт	шт	шт'	ш'т	ш'т'
шш	шш	шш'	ш'ш	ш'ш'

Finally, diphones database of Kazakh language totaled about 1000 diphones.

3. Phonetic transkriptor

For the development of phonetic transkriptor were investigated orthoepic rules of the Kazakh language. For the conveniance of the reader in this text, the rules are divided into groups that are numbered:

1. In the Kazakh language, if the word begins with vowel «е», then in front of it pronunciation is heard as «й», if the word starts with the vowel «о», «ө», then the pronunciation in front of them formed a brief insert «у» for example, «ет» – «йет», «он» – «уон», «өнер» – «уөнер».

2. If a word begins with a consonant «р» or «л», then the pronunciation of these sounds could be heard before the vowel «ы», «і», depending on the hardness or softness of consonants here «r», «l» means soft analogs "«р» and «л». For example, «рас» – «ырас», «рет» – «ірет», «лас» – «ылас», «лезде» – «ілезде».

3. When pronouncing borrowed sound «ю» as part of word is heard «йұу», «йұу»,, depending on the hardness or softness of the other vowels in syllables. For example: «қою» – «қойұу», «түю» – «түйұу»;;

4. When pronouncing borrowed sound «я» as a part of a word is heard «йа», «йә», depending on the hardness or softness of the other vowels in syllables. For example: «аян» – «айан», «әлия» – «әлійә»;

5. When pronouncing borrowed sound «и» in a consisting of words heard «ый», «ій», depending on the hardness or softness of the other vowels in syllables. For example, «ине» – «ійне», «жина» – «жыйна». If before or after «и» go according to «қ», «ғ» with descender, that the pronunciation of the sound «и» always heard «ый». For example, «қиын» – «қыйын», «қиғаш» – «қыйғаш».

6. When the pronunciation of the diphthong «у» as a part of a word is heard «ұу», «үу», depending on the hardness or softness of the other vowels in syllables. For example, «туыс» – «тұуыс», «күту» – «күтүу».

7. The Vowels «ұ», «ү», «о», «ө» at the beginning or the first syllable of the word in the pronunciation change in the next syllable vowel sounds «ы», «і» on the vowels «ұ», «ү» respectively. For example, «қолтық» – «қолтұқ», «құлын» – «құлұн», «күлкі» – «күлкү», «көлік» – «көлүк»;

8. Vowels «ү», «ө» in the beginning or in the first syllable of the word during the pronunciation changes in the following syllables vowel «е» to the next vowel «ө», for example, «үлкен» – «үлкөн», «өнер» – «өнөр».

9. Vowels «ә», «ү», «і» in the beginning or in the first syllable of the word during the pronunciation changes in the following syllables vowel «а» to its allophone «ә», for example, «ләззат» – «ләззәт», «діндар» – «діндәр».

10. If in a word sounds «с» and «ш», «с» and «ж» or «з» and «ш» meet in succession, then instead of them is pronounced the sound of double «шш». Also, instead of borrowed sound «щ» is pronounced «шш». For example: «досжан» – «дошшан», «басшы – башшы», «сөзшең – сөшшөң», «көшсең»-«көшшөң», «ащы»-«ашшы».

11. If in a word after sounds «з» and «ж» meet in succession, instead of them pronounce dual sound «жж», if sounds «з» and «с» meet in succession, instead of them pronounce dual sound «сс». For example: «бозжорға» – «божжорға», «азсыну – ассынұу».

12. If in a word after sound «н» встречается «б» или «п», then the pronunciation of sound «н» replaced to «м». For example: «мінбер – мімбер», «ойынпаз» – «ойымпаз».

13. If in a word after sound «н» meet «г», «ғ», «к» or «қ» then pronounce of «н» replaced to «ң». For example: «түнгі» – «түңгү», «қашанғы» – «қашаңғы», «зиянкес» – «зыйаңкес», «сәнқой» – «сәңқой».

14. When pronouncing the word in the composition of sound combinations мл, ғн, ғл between two sounds is formed a brief insertion of vowels «ы», «і», depending on the hardness and softness corresponding syllable. For example, «мемлекет» – «мемілекет», «бағлан» – «бағылан», «яғни» – «йағыный».

15. Uncombinable sounds found in many compound words are replaced by the pronunciation sound. For example, «шашбау» – «шашпау», «атбегі»-«атпегі», «атжалман» – «атшалман», «Көпбосын» – «Көппосұн», «түпдерек» – «түбдөрөк», «көпжиын»– «көбжыйын», «көпмүше – «көбмүшө», «түпнегіз» – «түбнегіз», «тасбауыр» – «таспаұуұр» и т.д.

Transkriptor implemented as a program that replaces some other characters in accordance with the rules contained in the control file. Rules are written in accordance with the each item of orthoepic rules of Kazakh language:

1. #е=йе, #о=уо, #ө=уө;

2. #л^а=ыл^а, #л^о=ыл^о, #л^ұ=ыл^ұ, #л^ә=іл^ә, #л^ү=іл^ү, #л^е=іл^е, #л^і=іл^і, р^а=ыр^а, #р^о=ыр^о, #р^ұ=ыр^ұ, #р^ә=ір^ә, #р^ү=ір^ү, #р^е=ір^е, #р^і=ір^і;

3. аю=айұу, ою=ойұу, ұю=ұйұу, ыю=ыйұу, үю=үйүу, ею=ейүу, қию=қыйұу, #тию#=тійүу, кию=кійүу, #сию#=сыйұу, #жию#=жыйұу, а^ию=а^ыйұу, о^ию=о^ұйұу, ұ^ию=ұ^ұйұу, ы^ию=ы^ыйұу, ә^ию=ә^ійүу, ө^ию=ө^үйүу, ү^ию=ү^үйүу, і^ию=і^ійүу, е^ию=е^ійүу;

4. ая=айа, оя=ойа, ұя=ұйа, ыя=ыйа, қия=қыйа, #сия=сыйа, #жия=жыйа, #мия=мыйа, #зия=зыйа, а^ия=а^ыйа, о^ия=о^ұйа, ұ^ия=ұ^ұйа, ы^ия=ы^ыйа, ә^ия=ә^ійә, ү^ия=ү^үйә, ия^а=ыйа^а;

5. #ми=мый, #жи=жый, а^и=а^ый, о^и=о^ый, ұ^и=ұ^ый, ы^и=ы^ый, ә^и=ә^ій, ө^и=ө^ій, ү^и=ү^ій, і^и=і^ій, е^и=е^ій, и^а=ый^а, и^о=ый^о, и^ұ=ый^ұ, и^ы=ый^ы, и^ә=ій^ә, и^ө=ій^ө, и^ү=ій^ү, и^і=ій^і, и^е=ій^е, қи=қый, ғи=ғый, иқ=ыйқ, иғ=ыйғ;

6. а^у=а^ұу, о^у=о^ұу, ұ^у=ұ^ұу, ы^у=ы^ұу, ә^у=ә^үу, ө^у=ө^үу, ү^у=ү^үу, і^у=і^үу, е^у=е^үу, у^а=ұу^а, у^о=ұу^о, у^ұ=ұу^ұ, у^ы=ұу^ы, у^ә=үу^ә, у^ө=үу^ө, у^ү=үу^ү, у^і=үу^і, у^е=үу^е;

7. о^ы=о^ұ, ұ^ы=ұ^ұ, ө^і=ө^ү, ү^і=ү^ү;

8. ө^е=ө^ө, ү^е=ү^ө;

9. і^а=і^ә, ү^а=ү^ә, ө^а=ө^ә;

10. сш=шш, сж=шш, зш=шш, шс=шш, щ=шш;

11. зж=жж, зс=сс;

12. нб=мб, нп=мп;

13. нг=ңг, нғ=ңғ, нк=ңк нқ=ңқ;

14. мл = міл, ғн=ғын, ғл=ғыл;

15. шб=шп, тб=тп, тж=тш, пб=пп, пд=бд, пж=бж, пм=бм,

пн=бн, сб=сп, сд=ст, қб=кп, қг=кг, қд=ғд, қж=ғж, қз=ғз, қм=ғым, қн=ғын, кб=кп, кг=кк, кд=кт, кж=гж, кз=гз, км=гм, кн=гн, зк=зг, зп=зб, зт=ст, қл=ғыл.

Each substitution rule is composed of two parts separated by a sign «=». From the left of this sign are original alphabetic character for word recording, on the right – the characters that should be replaced in the transcription.

For transcription of the given word consistently searches next occurrence of the left part of rule in it. If any of it detected, then instead of it, inserted the right part of the rule.

As a transcription symbols for vowels used mainly relevant Kazakh letters. Solid consonants are transcribed as Kazakh letters and the relevant soft consonants with the analogous Latin letters.

«#»means the beginning or end of word, depending on the location of: if«#» standing in front of characters, then it is the beginning of word; if «#» stands after the characters, it's the end.

«^»means any characters in any number between two sounds.

Each substitution rule is composed of two parts separated by a sign «=». From the left of this sign are original alphabetic character recording word on the right – the characters that should be replaced in the transcription.

For transcription of given word consistently searches next occurrence of the left part of rule in it. If any of it is detected, then it is inserted along the right side of the rule.

It is recommended that in the control file of these groups in numerical order, without changing the order of the rules in groups, because the order of substitutions is obviously important.

Also orthoepic rules were included to transkriptor rules defining the softness and labial consonants:

1. әб=әb, әг=әg, әд=әd, әж=әv, әз=әz, әй=әj, әк=әk, әл=әl, әм=әm, ән=әn, әң=әq, әп=әf, әр=әr, әс=әs, әт=әt, әу=әu, әш=әw, еб=eb, ег=eg, ед=ed, еж=ev, ез=ez, ей=ej, ек=ek, ел=el, ем=em, ен=en, ең=eq, еп=ef, ер=er, ес=es, ет=et, еу=eu, еш=ew, іб=ib, іг=ig, ід=id, іж=iv, із=iz, ій=ij, ік=ik, іл=il, ім=im, ін=in, ің=iq, іп=if, ір=ir, іс=is, іт=it, іш=iw, бә=bə, гә=gə, дә=də, жә=və, зә=zə, йә=jə, кә=kə, лә=lə, мә=mə, нә=nə, ңә=qə, пә=fə, рә=rə, сә=sə, тә=tə, уә=uə, шә=wə, бе=be, ге=ge, де=de, же=ve, зе=ze, йе=je, ке=ke, ле=le, ме=me, не=ne, ңе=qe, пе=fe, ре=re, се=se, те=te, ше=we, бі=bi, гі=gi, ді=di,

жі=vi, зі=zi, йі=ji, кі=ki, лі=li, мі=mi, ні=ni, ңі=qi, пі=fi, рі=ri, сі=si, ті=ti, ші=wi.

2. өб=өb, өг=өg, өд=өd, өж=өv, өз=өz, өй=өj, өк=өk, өл=өl, өм=өm, өн=өn, өң=өq, өп=өf, өр=өr, өс=өs, өт=өt, өу=өu, өш=өw, үб=үb, үг=үg, үд=үd, үж=үv, үз=үz, үй=үj, үк=үk, үл=үl, үм=үm, үн=үn, үң=үq, үп=үf, үр=үr, үс=үs, үт=үt, үу=үu, үш=үw, бө=bө, гө=gө, дө=dө, жө=vө, зө=zө, йө=jө, кө=kө, лө=lө, мө=mө, нө=nө, ңө=qө, пө=fө, рө=rө, сө=sө, тө=tө, уө=uө, шө=wө, бү=bү, гү=gү, дү=dү, жү=vү, зү=zү, йү=jү, кү=kү, лү=lү, мү=mү, нү=nү, ңү=qү, пү=fү, рү=rү, сү=sү, тү=tү, уү=uү шү=wү;

Let us explain marks used in the replacement rules. Latin characters in a group of 16 means that the sound is soft and non-labial, 17 in group "2" after the consonant means that the consonant – solid lip and in the group of 18 the number "3" after the consonant means that the consonant – soft labial.

In general phonetic transkriptor was about 400 rules.

4. Synthesis of etalon words

Standards of recognized words by the dictionary form of diphone standards, full base which in the volume of three thousand created for each speaker in advance [9]. Note that the creation of such a database in the future eliminates the need to establish any standards of voice.

Under diphones corresponding interphoneme transition within a word, we mean portion of standard length: 3 windows in 368 samples left of the label between the sounds and 3 of the same window to the right of the same label. Standard diphones – a set of 6 of the corresponding vectors. In addition, we use a window portion 3 to beginning of the word and the portion 3 of the window to the word conditionally calling their respective initial and final half diphones word (the transition from silence to speech and vice versa). All vectors in samples of diphones, play the role of code vectors and form a codebook B All samples of diphones and all code vectors are numbered.

Every word of the dictionary automatically transcribed, transcription construct a chain names of diphones. Each of them is replaced by the corresponding standard diphone. Obtained chain of vectors forms a standard word [9].

5. Testing

As a result of work it was built system, to recognize the word on the standards synthesized from diphones.

In testing of this system took part 5 speakers. For each of them have been created their own diphone database of two types: the base of 500 diphones in which every sound combination has only one analog and full base consisting of 1000 diphones in which consonants sound combinations have 4 options. After creating a diphone database, speakers uttered 50 words twice: to recognize words on the basis of incomplete diphone database and to recognize words on the basis of full diphone database. As a result of the recognition of words on basis of full diphone database reliably turned by approximately 15%.

Full results are shown in a Table 6.

Table 4

Results of word recognition Page margins

Sound	Whole base	Not whole base
Speaker 1	95,4%	80,1%
Speaker 2	94,8%	78,8%
Speaker 3	95,5%	79,4%
Speaker 4	93,5%	77,5%
Speaker 5	94,2%	75,5%

Thus, it appeared that usage of the expanded base diphones more effective and reliable.

6. Conclusion

What are the results obtained? Firstly, we were able to detect extremely large vocabularies, as the automatic generation of standards facilitates learning system. DTW algorithm is quite reliable for this purpose. We believe that independence from speaker could be achieved by averaging the standards. But even as it is depends from speaker, creating diphone database will take a maximum of 2–3 hours.

The most difficult in this technology is the transition to continuous speech, as it is difficult to determine the boundaries of words in continuous speech. Then, instead of the usual vocabulary need a text body with all sorts of suggestions and phrases. You can recognize a

combination of phrases as whole words, but such combinations will be many. Therefore, the effective usage of such a system for a particular subject area.

But once we have made a step in the direction of large dictionaries, this problem might be solved with hard work.

REFERENCES

1. Raut, C.K.Bayesian discriminative adaptation for speech recognition. – Acoustics, Speech and Signal Processing, IEEE International Conference on Eng. Dept., Cambridge Univ., Cambridge, 19–24 April 2009, Page(s): 4361 – 4364.

2. Lawrence, R. A tutorial on Hidden Markov Models and selected applications in speech recognition. – Proceedings of the IEEE 77 (2), 1989 pp. 257–286.

3. Mohamad Adnan Al-Alaoui, Lina Al-Kanj, Jimmy Azar, and Elias Yaacoub, Speech Recognition using Artificial Neural Networks and Hidden Markov Models, IEEE Multidisciplinary Engineering Education Magazine, Vol. 3, No. 3, September 2008.

4. Винцюк, Т.К. Анализ, распознавание и интерпретация речевых сигналов. Киев, Наук. думка, 1987, 262 стр.

5. Najkar, N., Razzazi, F., Sameti, H. An evolutionary decoding method for HMM-based continuous speech recognition systems using particle swarm optimization, Pattern Analysis And Applications, vol. 17, no. 2, pp. 327–339, 2014.

6. Frikha, M., Hamida A.B. A Comparitive Survey of ANN and Hybrid HMM/ANN Architectures for Robust Speech Recognition American Journal of Intelligent Systems 2012 2(1): 1–8

7. Hosom, J.P., Cole R., Fanty, M. Speech Recognition Using Neural Networks at the Center for Spoken Language Understanding. – Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, 1999.

8. Limkar, M., Rama, R., Vidya, S. Isolated Digit Recognition Using MFCC AND DTW. – International Journal on Advanced Electrical and Electronics Engineering, (IJAEEE), Vol.1, Issue-1, 2012, pp. 59–64.

9. Шелепов, В.Ю., Ниценко А., Дорохина, Г.В., Карабалаева, М.Х., Бурибаева, А.К. О распознавании речи на основе межфонемных переходов. – Вестник, Специальный выпуск, Евразийский национальный университет им. Л.Н. Гумилева, 2012, стр. 436–440.

10. Есенбаев, Ж., Махамбетов, О., Карабалаева, М. Текстовый корпус казахского языка – материалы международной научно-практической кон-

ференции «Современное казахское языкознание: актуальные вопросы прикладной лингвистики», 2012, стр. 61–66.

11. Raut, C.K.Bayesian discriminative adaptation for speech recognition. – Acoustics, Speech and Signal Processing, IEEE International Conference on Eng. Dept., Cambridge Univ., Cambridge, 19–24 April 2009, Page(s): 4361 – 4364.

12. Lawrence, R. A tutorial on Hidden Markov Models and selected applications in speech recognition. – Proceedings of the IEEE 77 (2), 1989 pp. 257–286.

13. Mohamad Adnan Al-Alaoui, Lina Al-Kanj, Jimmy Azar, and Elias Yaacoub, Speech Recognition using Artificial Neural Networks and Hidden Markov Models, IEEE Multidisciplinary Engineering Education Magazine, Vol. 3, No. 3, September 2008.

14. Винцюк, Т.К. Анализ, распознавание и интерпретация речевых сигналов. Киев, Наук. думка, 1987, 262 стр.

15. Najkar, N., Razzazi, F., Sameti, H. An evolutionary decoding method for HMM-based continuous speech recognition systems using particle swarm optimization, Pattern Analysis And Applications, vol. 17, no. 2, pp. 327–339, 2014.

16. Frikha, M., Hamida A.B. A Comparitive Survey of ANN and Hybrid HMM/ANN Architectures for Robust Speech Recognition American Journal of Intelligent Systems 2012 2(1): 1–8.

17. Hosom, J.P., Cole R., Fanty, M. Speech Recognition Using Neural Networks at the Center for Spoken Language Understanding. – Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, 1999.

18. Limkar, M., Rama, R., Vidya, S. Isolated Digit Recognition Using MFCC AND DTW. – International Journal on Advanced Electrical and Electronics Engineering, (IJAEEE), Vol.1, Issue-1, 2012, pp. 59–64.

19. Шелепов, В.Ю., Ниценко А., Дорохина, Г.В., Карабалаева, М.Х., Бурибаева, А.К. О распознавании речи на основе межфонемных переходов. – Вестник, Специальный выпуск, Евразийский национальный университет им. Л.Н. Гумилева, 2012, стр. 436–440.

20. Есенбаев, Ж., Махамбетов, О., Карабалаева, М.Текстовый корпус казахского языка – материалы международной научно-практической конференции «Современное казахское языкознание: актуальные вопросы прикладной лингвистики», 2012, стр. 61–66.

FORMATION OF STANDARDS PHONEMES BASED ON FAST CONTINUOUS WAVELET TRANSFORM OF THE SPEECH SIGNAL

Valeriy Zheltov[2], Pavel Zheltov[3], Vladimir Semenov[4], Alexander Shurbin[5]

[2]Chuvash State University, Cheboksary, Russia (428015, Cheboksary, Russia, Moskovskiy prospect 15), e-mail: zheltov42@mail.ru[1]
[3]Chuvash State University, Cheboksary, Russia (428015, Cheboksary, Russia, Moskovskiy prospect 15), e-mail: chnk@mail.ru
[4]Chuvash State University, Cheboksary, Russia (428015, Cheboksary, Russia, Moskovskiy prospect 15), e-mail: Alexgubm@gmail.com
[5]Chuvash State University, Cheboksary, Russia (428015, Cheboksary, Russia, Moskovskiy prospect 15), e-mail: Alexgubm@gmail.com

The speech signal is an example of a non-stationary process in which informative is the fact of changing the time-frequency characteristics. Studies show that the construction of appropriate algorithms for analysis of speech signals are models based on the frequency representation of the segments of the wavelet spectrum of the speech signal. For calculation the wavelet spectrum of a speech signal is used the formula of continuous wavelet transformation. Wavelet analysis of the speech signal indicates that the vowel phonemes have maximum energy at the average values of the scale factor. This pattern is observed when repeated many times, and does not depend on random factors.

Речевой сигнал является примером нестационарного процесса, в котором информативным является сам факт изменения его частотно-временных характеристик. Для выполнения анализа таких процессов требуются базисные функции, обладающие способностью выявлять в анализируемом сигнале как его частотные, так и временные характеристики. Такими базисными функциями являются вейвлеты. Вейвлеты, как средство многомасштабного анализа, позволяют выделять одновременно как основные характеристики сигнала, так и короткоживущие высокочастотные явления в речевом сигнале [1, 2]. Получение дополнительной информации в разных масштабах времени и разрешения сигнала может улучшить точность распознавания речи. Для вычисления вейвлет-спектра

[1] Работа выполнена при поддержке РФФИ, проект № 14-07-00143 а.

речевого сигнала используется формула непрерывного вейвлет-преобразования.

$$W(a,b)=\frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} S(t)\ \psi\left(\frac{t-b}{a}\right)dt.$$

Так как при распознавании речи необходима большая скорость преобразования, непрерывное вейвлет-преобразование вычисляется в частотной области с использованием быстрого преобразования Фурье [4, 5].

Прежде чем применять вейвлет-преобразование для формирования эталонов фонем, вейвлет-спектры $W(a,b)$ речевого сигнала проверялись на слух. Использовался вейвлет на основе второй производной функции Гаусса. Для прослушивания вейвлет-спектров $W(a,b)$ применялся алгоритм, который включает следующие шаги.

1. Находится максимум функции $W(a,b)$ в окне наблюдения.

2. Функция $W(a,b)$ нормируется.

3. Нормированная функция умножается на число, приемлемое по громкости.

4. Полученная функция округляется до целых чисел и подается на прослушивание.

Прослушивание вейвлет-спектров $W(1,b)$, $W(2,b)$, $W(4,b)$, $W(6,b)$, $W(8,b)$ показывает, что они звучат почти идентично исследуемому речевому сигналу $S(t)$. Например, функция $W(3,b)$ фонемы *е* звучит как фонема *е* с другим тембром. При увеличении масштабного коэффициента a более 20 для выборки 32768 отсчетов каждая фонема теряет свою индивидуальность и звучит как фонема *у* [3].

Для формирования эталонов фонем применяется следующий алгоритм. Вычисляются вейвлет-коэффициенты $W(1,b)$, $W(2,b)$ слов, где b меняется от 1 до 32768. Полученные вейвлет-коэффициенты (функции) $W(1,b)$, $W(2,b)$ и $S(t)$, предварительно очищенный от низкочастотных составляющих, разбиваются на сегменты фиксированной длительности (n = 128), что соответствует 16 мс. Количество сегментов равно 256. Длительность сегмента не меньше длительности произношения фонем, но превышает максимально возможный период основного тона фонем.

В каждом сегменте вычисляется Фурье спектр от вейвлет-коэффициентов $W(1,b)$, $W(2,b)$ и $S(t)$. Таким образом, математической моделью речевого сигнала в сегменте является:

$$d(n) = \frac{1}{M}\sum_{k=0}^{M-1} W(a,k)\cos\left(\frac{2\pi nk}{M}\right),$$

$$e(n) = \frac{1}{M}\sum_{k=0}^{M-1} W(a,k)\sin\left(\frac{2\pi nk}{M}\right).$$

По формуле

$$F(i) = d^2(i) + e^2(i)$$

вычисляется Фурье-спектр функций $W(1,b)$, $W(2,b)$ и $S(t)$ фонем русского и чувашского алфавита. На рис. 1 и 2 приведены Фурье-спектры сегментов функции $W(1,b)$ фонем а и э.

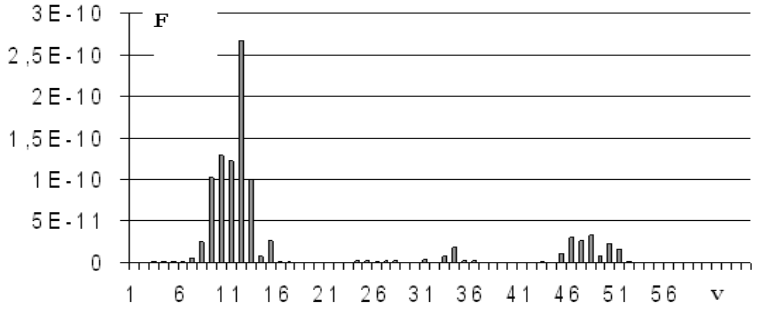


Рис. 1. Фурье-спектр функции $W(1, b)$ фонемы *а*

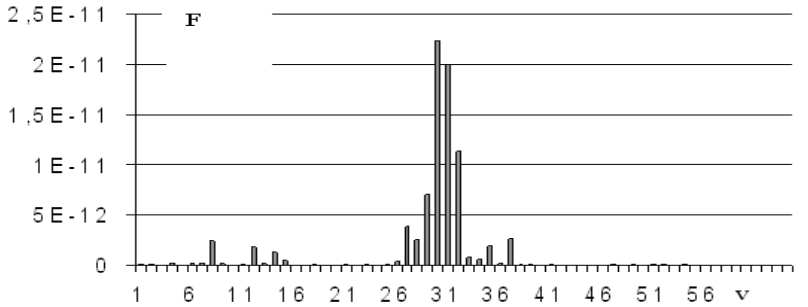


Рис. 2. Фурье-спектр функции $W(1,b)$ фонемы э

Для фонем русского и чувашского алфавита создана база данных с набором характерных частот сегментов функций $W(1,b)$, $W(2,b)$ и

$S(t)$. Также в качестве характерного признака фонем используется количество переходов через ноль функций $W(1,b)$, $W(2,b)$ в сегменте, принятые как определенные частоты. Несколько вариантов характерных частот, полученные путем многократного произношения русских и чувашских слов на этапе обучения, используются как эталоны фонем для распознавания речи. Эталоны фонем разбивается на четыре группы, так как при вейвлет-преобразовании с различными масштабными коэффициентами a, гласные и согласные фонемы имеют различные значения по величине. В свою очередь, шипящие фонемы при малых значения масштабного коэффициента a имеют сравнимые с гласными фонемами вейвлет-коэффициенты, то есть на месте выделения согласной фонемы образуются максимумы, которые легко выделить. Взрывные звуки выделяются при больших масштабных коэффициентах a и так же имеют характерный провал по величине вейвлет-коэффициентов перед собой.

ЛИТЕРАТУРА

1. Дремин И.Л. Иванов О.В., Нечитайло В.А. Вейвлеты и их использование // УФН. 2001. Т. 171. № 5. С. 465–501.

2. Желтов П.В., Семенов В.И. Вейвлет-анализ акустического сигнала КГТУ им. А.Н. Туполева// Вестн. КГТУ. 2008. Вып. 4.

3. Желтов П.В., Семенов В.И. Вейвлет-преобразование акустического сигнала/ КГТУ им. А.Н. Туполева. Казань, 2008. 102 с.

4. Семенов В.И. Свидетельство об официальной регистрации программы для ЭВМ № 2007615024. Непрерывное быстрое вейвлет-преобразование / Зарег. в Реестре программ для ЭВМ 4 декабря 2007 г.

5. Семенов В.И., Желтов П.В. Патент на изобретение № 2403628 РФ, МПК G10L 15/10. Способ распознавания ключевых слов в слитной речи. Опубл. 10.11.2010 Бюл. №31.

DEVELOPMENT OF THE CHUVASH SPEECH SYNTHESIS SYSTEM BASED ON A COMBINED DIPHONES DATABASE

Evgeniy Sergeev, Andrey Skvortsov[1]

[1]Chuvash State University
Cheboksary, Chuvashia, Russia

As a result of this work will be solved tasks traditionally assigned to artificial intelligence, especially tasks related to knowledge representation and processing of natural language texts will be developed and implemented algorithms Chuvash speech synthesizer and the estimation of their complexity. The project involves a major speech research, the object of study – a natural language, the actual problem – the presentation and recognition of (interpretation of) the meaning of linguistic expressions.

Атуальность и современное состояние исследований по данной научной проблеме

Создание систем синтеза речи является одним из наиболее актуальных направлений развития современных компьютерных технологий. В зарубежной литературе для обозначения систем компьютерного синтеза речи принято выражение text-to-speech (TTS). Первые разработки в области синтеза речи относятся к 1930-м годам и связаны с применением разработок в телефонии компанией Bell Labs. Однако, проблемы, выявленные на ранних этапах данной научной задачи, не решены и сегодня: при реализации синтеза речи по правилам, автоматическом формировании речевого сигнала по транскрипции или орфографическому представлению содержания сообщения возникают проблемы качества синтезируемого сообщения, часто из речевых единиц складывается трудно понимаемая речь. Стоит отметить, что в последние 10 лет качество систем синтеза речи было существенно учучшено за счет расширения используемых фонетических словарей и моделирования супрасегментных единиц. Технологии TTS используются в информационно-справочных системах с речевым доступом, в оперативном формировании публичных объявлений (на вокзалах, в аэропортах и пр.), для информирования операторов о технических процессах, в качестве адаптированного программного обеспечения для тех у кого ослаблено или полностью отсутствует зрение, в последнее время широко внедряются в образование.

Сегодня разработкой систем синтеза речи занимаются сотни компаний и исследовательских групп, среди которых следует назвать Acapela Group, Nuance,The Centre for Speech Technology Research (CSTR) эдинбургского университета и др.

В рамках речевых исследований можно выделить следующие фундаментальные проблемы, на решение которых направлен проект: нейрофизиология управления артикуляцией, обучение языку, компенсация и адаптация к помехам артикуляции, связь между артикуляцией и акустикой, механизмы восприятия, распознавания и понимания речи человеком.

Конкретной фундаментальной проблемой является разработка и реализация программно-информационной оболочки синтезатора чувашской речи и создание необходимой для работы системы акустических баз данных связи с артикуляцией.

Российский рынок представлен единичными разработками систем синтеза речи. В результате анализа современного состояния данной научной проблемы, прямых аналогов не выявлено. Косвенным аналогом является синтезатор татарской речи НИИ «Прикладная семиотика» Академии наук РТ.

Предлагаемые методы и подходы решения поставленной задачи:

1. Анализ систем и технологий в области распознавания и синтеза речи.

2. Исследования по определению значимости слоговой границы в фонетическом или фонологическом отношениях для языка с малой контекстной зависимостью.

3. Разработка акустической базы данных для синтезатора чувашской речи.

4. Программная реализация транскрипционного анализатора, предназначенного для преобразования входного орфографического текста в размеченный фонемный текст.

5. Программная реализация генерации позиционных и комбинаторных звуковых единиц (дифонов

6. Разработка программного комплекса синтезара речи с поддержкой чувашского и русского языка.

7. Тестирование программного комплекса.

Цель работы – создание дифонного синтезатора речи. Данный продукт так же должен содержать базу данных дифонов для жен-

ского и мужского голосов. Программа должна иметь предельно ясный интерфейс и должен быть ориентирован на людей различных возрастов и имеющих различные навыки пользования персональным компьютером.

Разрабатываемый проект должен удовлетворять следующим требованиям:

- Иметь возможность синтеза чувашского текста как путем загрузки текста в редактор синтезатора, так и путем простого набора в редакторе.
- Синтезатор речи должен воспроизвести любое введенное чувашское слово.
- Для синтеза речи требуется нарезать дифонную базу данных женского и мужского голосов.

В основу решения проблемы положен метод параметрического представления. С целью уменьшения требуемой памяти для хранения базы звуков и обеспечения необходимой гибкости. Этот метод синтеза экономичнее волнового, так как требует значительно меньшего объема памяти, но при этом он требует больше вычислений, чтобы воспроизвести исходный речевой сигнал. Для синтеза используются единицы речи – дифоны. Преимущества этого метода: гибкость, немного памяти для хранения исходного материала, сохранение индивидуальных характеристик диктора.

Входной орфографический текст подвергается ряду последовательных обработок с помощью специальных процессоров. Транскрипционный анализатор предназначен для преобразования входного орфографического текста в размеченный фонемный текст. Под разметкой подразумевается разбиение текста на отдельные элементы. Размеченный фонемный текст поступает на вход дифонного анализатора. Задача дифонного анализатора заключается в генерации позиционных и комбинаторных дифонов. Акустический процессор на основе информации о том, какие дифоны необходимо использовать, генерирует речевой сигнал путем компиляции отрезков естественных звуковых волн соответствующих дифонов. Общая структура дифонного синтезатора представлена следующим образом (рис.1).

При решении задачи звучания синтезированной речи предусмотрены возможные искажения элементов компиляции в процессах их записи, воспроизведении и модификации.

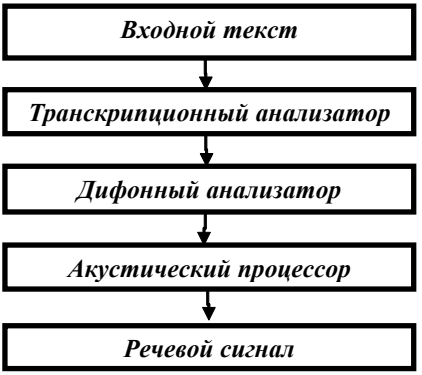


Рис. 1. Общая структура дифонного синтезатора речи

При решении задачи звучания синтезированной речи выполнены следующие требования:

1. Максимально полное использование при синтезе речи акустических, фонетических и просодических средств индивидуальности голоса и речи имитируемого диктора;

2. Минимально возможные искажения элементов компиляции в процессах их записи, воспроизведения и просодической модификации.

Выполнение данных требований позволило намного увеличить качество синтеза. В работе дается анализ правил изменения фонем в потоке речи, опираясь на которые созданы алгоритмы. Изложены основные принципы и основные алгоритмы, положенные в основу синтезатора. Приведены основные методы решения поставленной задачи, способы решения лингвистических особенностей чувашского языка.

СПИСОК ЛИТЕРАТУРЫ

1. Лобанов Б.М. «Ретроспективный обзор исследований и разработок Лаборатории распознавания и синтеза речи». Сб. «Автоматическое распознавание и синтез речи», ИТК НАН Беларуси, Минск, 2000.

2. Математическая лингвистика. Сб. переводов / под ред. Ю.А. Шрейдера, И.И. Ревзина, Д.Г. Лахути, В.К. Финна. – М.: Мир, 1964. – 144 с.

3. Сергеев Е.С., Пигачев П.В. ДИФОННЫЙ СИНТЕЗАТОР РЕЧИ. – Теоретические и прикладные аспекты современной науки. 2014. № 6-3. С. 114–116.

DESIGN AND CREATION OF SPEECH CORPORA FOR THE TATAR SPEECH RECOGNITION AND SYNTHESIS TASKS

Aidar Khusainov

Institute of Applied Semiotics of the Tatarstan Academy of Sciences
Kazan (Volga region) federal university, Kazan, Russia
khusainov.aidar@gmail.com

In this paper we describe our recent work of creation speech corpora for the Tatar language. Information from these corpora are used to create acoustic models for speech recognition and synthesis systems. We describe the procedure of corpora creation and their current characteristics. Programming tools and equipment used for recording and annotation are mentioned. We also present some future improvements that will allow to increase speech recognition and synthesis quality for the Tatar language.

1. Введение

Под термином корпус понимается информационная система, содержащая данные об одном или нескольких языках в электронной форме. При этом информация о языке может быть представлена в различных модальностях. Так, например, можно говорить о текстовых и речевых корпусах, хранящих информацию в текстовом и аудиоформатах.

Каждый корпус создаётся для решения конкретной задачи, что позволяет учитывать при их проектировании и создании специфику проводимых исследований. В данной статье представлены результаты работ по созданию корпусов татарского языка, которые создаются для использования в контексте речевых информационных технологий. Набор речевых записей требуется для создания статистических моделей звуков татарского языка.

Структура данной работы предполагает описание основных результатов создания корпусов татарского языка исходя из целей их использования. В пункте 2 приводится описание двух речевых корпусов, необходимых для функционирования систем распознавания речи; в пункте 3 описан речевой корпус, записанный для обучения системы синтеза татарской речи.

2. Корпусы для распознавания татарской речи

Характеристики корпуса, его структура, а также способ записи его элементов, во многом зависят от конечных целей использования корпуса. Разработка системы автоматического распознавания татарской речи подразумевает создание и использование моделей трёх уровней описания языка:

- Акустический уровень: модели произношения акустических единиц языка (фонем, дифонов и т.д.).
- Лексический уровень: словарь используемых слов, фонетические транскрипции слов.
- Уровень языковой модели: описываются правила употребления слов языка.

Исходя из этой классификации, появляется необходимость обеспечить модели всех 3 уровней исходной информацией, на основе которой могло бы осуществляться их обучение. Проблема создания словарей и способа построения фонетических транскрипций слов (осуществления графем-фонемных преобразований) для татарского языка была решена в [1]. В качестве текстового корпуса используется корпус татарского языка [2]. А для создания моделей акустического уровня используются созданные речевые корпусы.

2.1. Речевой корпус одного диктора

Целью создания данного корпуса послужила необходимость тестирования экспериментальной версии системы распознавания фонем татарского языка. Большая часть речевых записей использовалась для обучения акустических моделей, оставшиеся – для проведения тестирования системы распознавания.

Диктором были произнесены наиболее частотные слова татарского языка. Анализ частотности проводился на основе имеющегося корпуса из 25 миллионов слов.

Программные средства

Запись осуществлялась с помощью программы SigRS, позволяющей осуществлять последовательную запись и хранение множества речевых фрагментов. Дополнительной возможностью является отображение спектрограммы записываемой речи в режиме реального времени, рисунок 1.

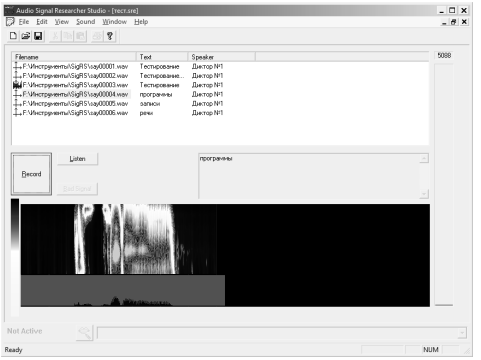


Рис. 1. Инструмент для записи речевых фрагментов SigRS

Оборудование и условия записи

Для записи использовался ноутбук HP Pavilion dv6 и аудиогарнитура Sennheiser PC330 G4ME. Запись осуществлялась в тихом офисном помещении.

Результат

Общее число записанных слов равняется 10788. Каждое слово было записано по одному разу в формате 22 kHz 16 bps PCM. Получившиеся записи были разделены на обучающую и тестовую часть в пропорциях 9 к 1. Основные характеристики корпуса представлены в таблице 1.

Таблица 1

Характеристики татарского речевого корпуса одного диктора

Параметр	Значение
Количество файлов	10788
Общая продолжительность	4:56:45
Количество файлов в обучающей части	9631
Продолжительность обучающей части	4:26:42
Количество файлов в тестовой части	1157
Продолжительность тестовой части	0:30:03

Построенная на основе данного корпуса экспериментальная система распознавания фонем показала 61% качество работы.

2.2. Многодикторный речевой корпус татарского языка

Целью является создание репрезентативного речевого корпуса татарского языка, содержащего читаемую и спонтанную речь множества дикторов. Корпус создаётся с целью использования в различных исследованиях: от экспериментальной фонетики до автоматического распознавания слитной татарской речи.

Основной мотивацией к созданию многодикторных корпусов звучащей речи служит большое разнообразие особенностей произношения. Чтобы система распознавания речи показывала устойчивые результаты при смене диктора она должна опираться в своей работе на акустические модели, аппроксимирующие максимальное количество особенностей произношения. Добиться этого можно за счет использования при обучении этих моделей речевых записей большого числа дикторов разного пола, возраста, диалекта, особенностей произношения.

Для того чтобы обеспечить выполнение требований по количеству дикторов, продолжительности записей, а также разнообразию используемой лексики, было принято решение о поэтапном создании корпуса татарской речи. Структура корпуса состоит из 3 основных частей, рисунок 2:

1. «Ядро» – часть корпуса, предназначенная для обеспечения максимально полного охвата фонем татарского языка на первом

Рис. 2. Структура многодикторного речевого корпуса татарского языка

этапе обучения моделей системы распознавания. Включает в себя небольшие по продолжительности записи большого количества дикторов.

2. Читаемая речь – часть корпуса, направленная на наполнение корпуса более продолжительными записями.

3. Спонтанная речь – часть корпуса, содержащая неподготовленную речь, заметно отличающуюся по своим характеристикам от читаемой.

Программные средства

При подготовке экспертами текстов для озвучивания использовались два основных критерия: необходимость обеспечить фонетическую полноту и разнообразие записанных фраз. Фонетическая полнота подразумевает под собой наличие в произнесенной фразе всех базовых акустических единиц языка (фонем). Под задачей максимизации фонетического разнообразия понимается увеличение количества разнообразных контекстов произнесения каждой фонемы.

Для выполнения данных требований был разработан и использован инструмент автоматического построения фонетической транскрипции татарских текстов и их статистического анализа. Эксперту были доступны гистограмма частотности фонем и информация о количестве 2- и 3-грамм, рисунок 3.

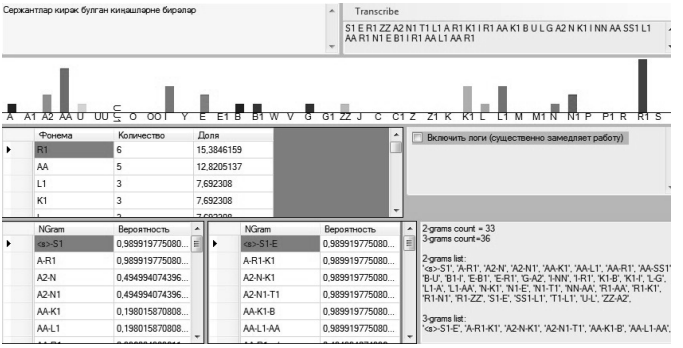


Рис. 3. Инструмент для анализа татарских текстов

Запись речевых фрагментов в подкорпусе «Ядро» осуществлялась в программе SigRS, в подкорпусе читаемой речи – в про-

грамме Audacity [3]. С учётом того, что в читаемой части корпуса производилась запись 30-минутного фрагмента текста, для дальнейшего использования было необходимо произвести процедуру сегментирования. Экспертами с помощью программы Transcriber AG [4] были выделены временные промежутки произнесения каждой интонационной группы.

Кроме того, была предложена модель разметки речевого корпуса, целью которой является ассоциация дополнительной информации о речевом сигнале, которая в дальнейшем может быть использована системой распознавания речи. Дополнительная информация вводится путём расстановки специальных тегов, которые можно разделить на несколько основных групп:

1. Теги используемого языка: *т* (татарский язык), *р* (русский), *англ* (английский), *т-рак* (татарский с русским акцентом), *р-так* (русский с татарским акцентом).

2. Теги обозначения особенностей произношения слов: *д* (диалекты), *ж* (жаргон), *о* (обрыв слова), *ог* (оговорка), *з* (заикание), *шеп* (шепот), *длин* (протяжное произнесение).

3. Теги обозначения шума: *кф* (кашель), *сф* (стук), *свф* (свист), *бф* (шелест бумаги), *мф* (музыка), *пнф* (пение), *хф* (одновременно несколько голосов), *смф* (смех), *плачф* (плач), *эф* (эхо), *вдф* (громкий вдох), *апф* (аплодисменты), *скрф* (скрип), *шмф* (шмыганье), *шф* (прочие шумы).

4. Теги для экстралингвизмов: *см* (смех), *плач* (плач), *к* (кашель), *вд* (вдох), *ч* (чавканье), *шм* (шмыганье), *аа*, *ее*, *мм*, *нн*, *ии*, *оо*, *уу* (заполненные паузы), *гм* (хмм).

Оборудование и условия записи

Для записи использовался ноутбук HP Pavilion dv6 с аудио-гарнитурой Sennheiser PC330 G4ME и ноутбук Lenovo IdeaPad U430p с аудио-гарнитурой Sennheiser G4ME ONE. Запись осуществлялась в тихих помещениях.

Результат

Основные характеристики корпуса, а также основные этапы его создания представлены в таблице 2.

Построенная на основе данного корпуса экспериментальная система распознавания отдельно произнесенных слов показала 92% качество работы, слитной татарской речи – 70% качество работы.

Таблица 2

Характеристики татарского многодикторного речевого корпуса и этапы его создания

Параметр	«Ядро»	Читаемая часть	Спонтанная часть	Весь корпус
Продолжительность	9:14:50	19:03:47	0	28:18:37
Количество дикторов	251	39	0	290
Средняя продолжительность записи 1 диктора	2:12	29:20	0	5:51
Этапы:				
1.1.Подготовка текстовых материалов (анализ фонетической полноты и разнообразия)	+	–	–	
1.2.Выбор текстовых материалов из различных источников	–	+	–	
2. Поиск дикторов	+	+	+	
3. Запись речевых фрагментов	+	+	+	
4. Стенографирование	–	–	+	
5. Сегментация	–	+	+	
6. Аннотирование	–	+	+	

3. Корпус для синтеза речи

Задача синтеза речи состоит в формировании аудиосигнала на основе фразы, представленной в текстовом виде. Большая часть подходов к синтезу речи основывается на конкатенативном подходе (дифонный синтез [5], Unit selection [6]). Исходной информацией в данных подходах служат выделенные из речевых записей акустические единицы. Начиная с 2002 года, набирает популярность альтернативный, параметрический, подход к синтезу речи, в котором базовыми элементами являются не выбранные фрагменты записи, а статистические модели звуков языка.

При разработке синтезатора татарской речи используется параметрический подход, при котором модели звуков строятся на основе скрытых Марковских моделей (HMM-based speech synthesis, HTS [7]).

Для обучения скрытых Марковских моделей необходимо наличие аннотированного речевого корпуса. Однако в отличие от корпуса для распознавания речи, который должен максимально полно описывать разнообразие голосов и способов произношения, корпус для синтеза речи должен отражать особенности произношения конкретного диктора, на основе которого будет строиться конечный синтезатор речи.

Оборудование и условия записи

Ещё одной особенностью создания корпуса для синтеза речи являются требования к качеству используемой аппаратуры и условиям записи. Для создании корпуса татарского языка был задействован профессиональный диктор, записи были сделаны в звукозаписывающей студии, в звуконепроницаемом помещении с использованием профессионального оборудования.

Результат

В записанных аудиофайлах экспертами были вручную размечены все интонационные группы, после чего была построена и использована система распознавания фонем. Таким образом, весь корпус был фонетически размечен. Для синтезатора речи были использованы скрытые Марковские модели, построенные отдельно для каждой из контекстно-зависимых фонем: в качестве контекста учитывались по две фонемы до и после анализируемой. Основные характеристики полученного корпуса представлены в таблице 3.

Таблица 3

Характеристики татарского речевого корпуса для синтеза речи

Параметр	Значение
Количество дикторов	1
Общая продолжительность	18:54:03
Частота дискретизации	44,1 kHz
Битрейт	24 bps

Планы

Для улучшения качества звучания синтезатора татарской речи необходимо увеличить количество анализируемых системой контекстов. На данный момент в качестве контекста используется информация исключительно о фонемах, звучащих до и после текущей.

Была предложена модель разметки корпуса, основные элементы которой можно представить следующим образом:

1. Уровень фонем: текущая фонема, две предшествующие, две последующие фонемы.

2. Уровень слогов: тип слога (V, VC, CV, CVC, VCC, CVCC); позиция фонемы в слоге; количество фонем в предыдущем, текущем, последующем слоге; номер текущего слога в слове; гласная в текущем слоге.

3. Уровень слов: часть речи, количество слогов для предыдущего, текущего, следующего слова; количество предшествующих и последующих слов во фразе.

4. Уровень фразы: количество слов/слогов в предыдущей, текущей, последующей фразе.

4. Заключение

Разработанные версии корпусов звучащей татарской речи позволяют строить акустические модели, необходимые для работы систем автоматического распознавания и синтеза речи.

Дальнейшее развитие корпусов предусматривает увеличение объёма речевого материала, включение записей со спонтанной речью, а также проведение работ по экспертной разметке корпусов с учётом разработанных моделей аннотирования.

ЛИТЕРАТУРА

[1] Khusainov, A. F., Suleymanov, Dz. Sh. Towards Automatic Speech Recognition for the Tatar Language / A. F. Khusainov, Dz. Sh. Suleymanov // Proc. of the 16th International Workshop on Computer Science and Information Technologies (CSIT'2014). (Sheffield, September 6–22, 2014). – Ufa: USATU, 2014. – P. 97–100.

[2] Сулейманов Д.Ш., Невзорова О.А., Галиева А.М., Гатиатуллин А.Р., Гильмуллин Р.А., Хакимов Б.Э. Размеченный корпус татарского языка

«Туган тел»: аспекты реализации // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2014. – Казань: Изд-во «Фэн» Академии наук РТ, 2014. – С. 88–93.

[3] Audacity [Электронный ресурс]. URL: http://audacityteam.org/about/.

[4] TranscriberAG [Электронный ресурс]. URL: http://transag.sourceforge.net/.

[5] Moulines, E., Charpentier, F. “Pitch-sinchronous waveform processing techniques for text-to-speech synthesis using diphones”. In Speech Communication, 9 (5/6), 1990, pp. 453–467.

[6] Sagisaka, Y. ATR v-talk speech synthesis system. In Proc. ICSLP-92, 1992, Banff, Canada.

[7] Yoshimura, T., Tokuda, K.. Masuko, T., Kobayashi, T., Kitamura, T. “Simultaneous modeling of spectrum, pitch and duration in HMM-basedspeech synthesis”. In Proc. Eurospeech, 1999, pp. 2347–2350.

СОДЕРЖАНИЕ

SECTION 1. **MACHINE TRANSLATION TECHNOLOGIES**

SECTION 2 . **LINGUISTIC SOFTWARE**

SECTION 3. **ELECTRONIC TEXT CORPORA**

SECTION 6. **SPEECH TECHNOLOGIES**

PROCEEDINGS
OF THE INTERNATIONAL CONFERENCE
“TURKIC LANGUAGES PROCESSING”
TurkLang-2015

Труды конференции
В авторской редакции

Подписано в печать 09.09.2015
Бумага офсетная. Формат 60x84 1/16.
Гарнитура «TimesNewRoman».
Печ. л. 30,5. Тираж 100 экз. Заказ №9/9.

Издательство Академии наук
Республики Татарстан
420111, г. Казань, ул. Баумана, 20
e-mail: izdat.anrt@yandex.ru

Отпечатано в ООИ «АРМАНД»
420126, г. Казань ул. Лаврентьева, д. 3