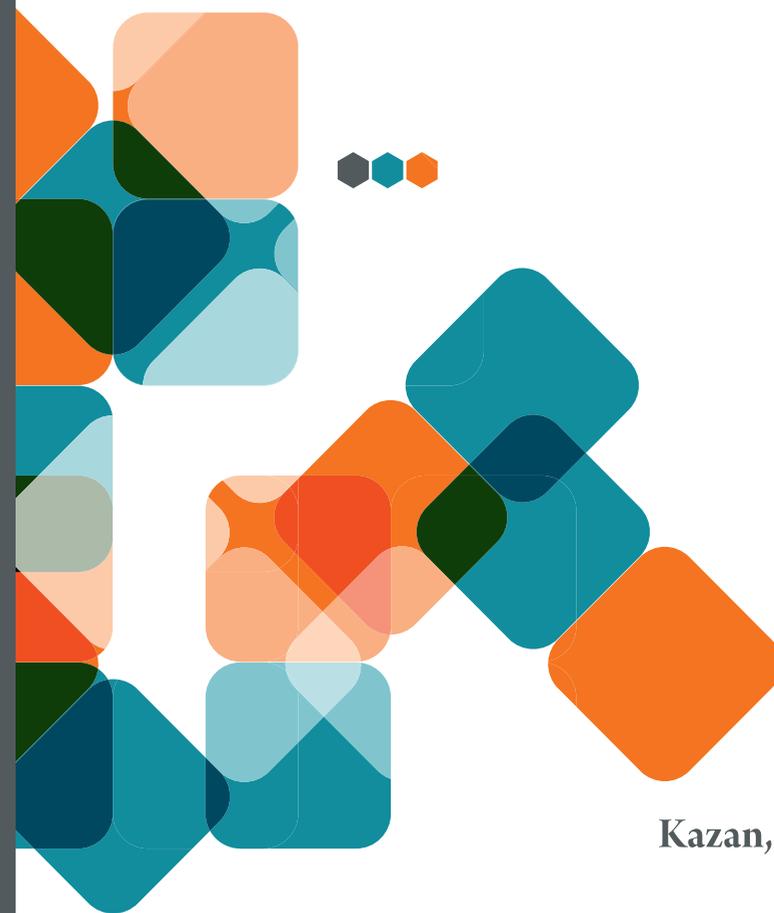


PROCEEDINGS
OF THE INTERNATIONAL CONFERENCE
“TURKIC LANGUAGES PROCESSING”

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 Information Technologies
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

© Academy of Sciences
of the Republic of Tatarstan Press, 2015

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 KFUA. 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 Ofrazier (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)

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

Abduali Balzhan¹, Akhmadieva Zhadyra², Zholdybekova Saule³,
Tukeyev Ualsher⁴, Rakhimova Diana⁵

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> <p><l></l><r><s n="m"/><s n="an"/><s n="sg"/><s
n="nom"/></r></p></e>
<e> <p><l></l><r><s n="m"/><s n="an"/><s n="sg"/><s
n="det"/></r></p></e>
<e> <p><l></l><r><s n="m"/><s n="an"/><s n="sg"/><s
n="ord"/></r></p></e>
...
</pardef>
<e><p><l>бip<s n="num"/></l><r>один<s n="num"/></r></
p><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><p><l></l><r><s n="m"/><s n="an"/><s n="sg"/><s
n="nom"/></r></p></e>
<e><p><l></l><r><s n="m"/><s n="an"/><s n="sg"/><s
n="det"/></r></p></e>
<e><p><l></l><r><s n="m"/><s n="an"/><s n="sg"/><s
n="ord"/></r></p></e>
<e> <p><l><s n="subst"/></l><r><s n="m"/><s n="an"/></
r></p></e>
<e> <p><l><s n="comp"/></l><r><s n="comp"/></r></p></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.