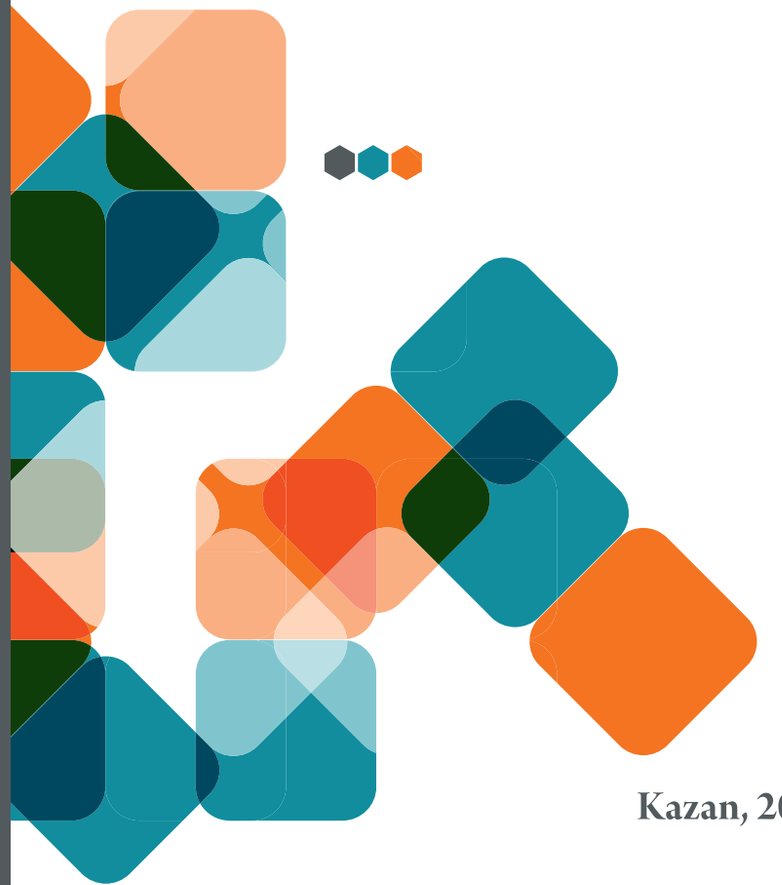
TurkLang • 2015

Tatarstan Academy of Sciences

PROCEEDINGS
OF THE INTERNATIONAL CONFERENCE
"TURKIC LANGUAGES PROCESSING"

TurkLang 2015

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



Kazan, 2015

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)

LEXICAL SELECTION RULES FOR KAZAKH-TO-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.

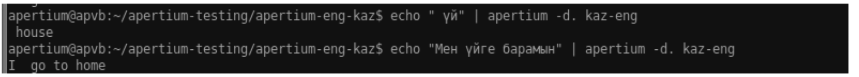


Fig. 1. Result of translating ambiguous noun.

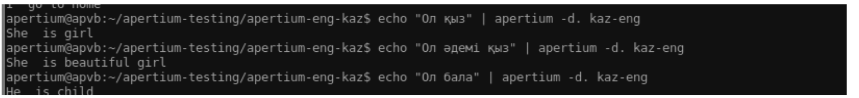


Fig. 2. Result of translating ambiguous pronoun.

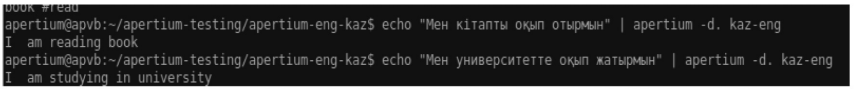


Fig. 3. Result of translating ambiguous verb

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.

СОДЕРЖАНИЕ

SECTION 1. **MACHINE TRANSLATION TECHNOLOGIES**

SECTION 2 . **LINGUISTIC SOFTWARE**

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