The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.38 e-ISSN: 2251-7545

Features of machine translation of different systemic languages using a Apertium platform

(with an example of English and Kazakh languages)

Shormakova Assem
Department of Information Systems
Kazakh National University,KazNU
Almaty,Kazakhstan
e-mail: assem007@mail.ru

Sundetova Aida
Department of Information Systems
Kazakh National University,KazNU
Almaty,Kazakhstan
e-mail: sun27aida@gmail.com

The goal of this article is to examine a grammatical and lexical problems, which we often face while translating English texts, and not giving any detailed statement of grammatical or lexical phenomenon. Only some sides of the given phenomenon are reviewed in the article, particularly the ones that represent linguistic-culturological interest in respect of translation from English to Kazakh language on a Apertium platform.

I. Introduction

By specificity of the reflected sides of the reality there are known similarities in the parts of speech system in Kazakh and English languages, in the same time they are characterized by a number of specific features, e.g.: vague borders of parts of speech in English language, rather than in Kazakh language. It is due to rather small development of English affixation, which leads to smaller formal placement of parts of speech [1]. There is no distinction in accentuation of the basic parts of speech, the greatest similarity is found in pronouns, in their lexical structure and their division into subclasses. There is a meaning difference in languages, for example, the definiteness of meaning is obligatory for English language, and is unessential for Kazakh language.

Though in English language the nouns lack any gender category, nevertheless the nouns that stand for animate objects, personal pronouns, and also, as an exception, some animals and subjects can have masculine or a feminine genders. The nouns that are related to the person (which answer the question Who is this? – Бұл кім?), happen to have a masculine gender (respective pronouns, he, his – ол, онікі) or a feminine gender (respective pronouns, she, her – ол, онікі). As we see from examples, in Kazakh language the gender category is completely absent, therefore, the same adjective, pronoun or an ordinal numeral can be translated into English differently, depends on the meaning of the sentence.

Translation is made on a Apertium platform. Apertium has arisen as the tool of machine translation within the OpenTrad project and has originally been intended for translation between related languages, however recently its possibilities have been expanded to cover less similar language pairs. To create the new system of machine translation there is a necessity to develop linguistic base (dictionaries, rules) in accurately specified XML formats.

Apertium is a platform of machine translation which is developed at financing from the governments of Spain and Catalonia at University of Alicante (Universitat d'Alacant). This is a free software which is published free of charge by developers according to GNU GPL conditions. [2]

The work essence is in solving a problem of sentence translation from English to Kazakh language. The word order both in Kazakh, and in English languages is very important, as it is constant. While translating to Kazakh language it is necessary to consider that adjectives and numerals before the noun don't change in number and a case.

Basically the word order in the sentence is the same as in English language. But the place of predicate is always at the end of the sentence.

Rules of translating such sentences are based on the latent models of Markov. Rules are described in the special dictionary and used as final converters for all lexical translations[3]. Hidden Markov model is used to translate different models of sentences when we don't have the same pattern. All pattern are saved in Regression Test.

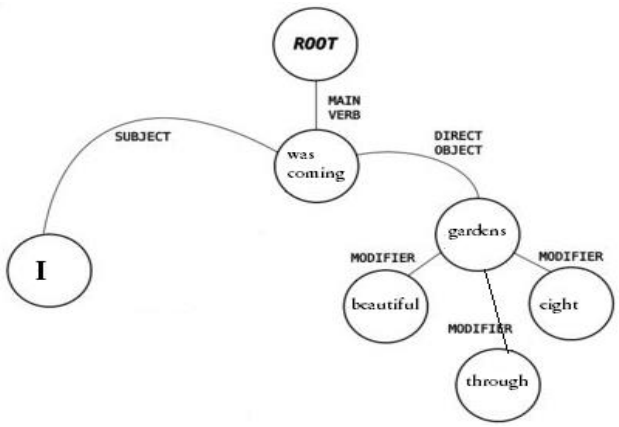
II. LEXICAL MODEL

The word order both in Kazakh, and in English languages is very essential, as it constant, e.g.: The second boy is my son – Екінші бала менің ұлым; It is raining – Жаңбыр жауып тұр; The blue sky is large – Аспан ашық еді. In Kazakh language the subject usually stays in the beginning of the simple narrative sentence, a predicate – always in the end, definition before a defined word, time adverb – in the beginning of the sentence. The same way there is a certain order of sentence parts in English language, where a subject and a predicate etc. have their own fixed places, e.g.: We live in Astana – Біз Астанада тұрамыз.
Here an example of analysis of a simple sentence: *Мен*

сегіз әдемі бақша арқылы келіп отырдым.

Transfer-based MT usually works as follows: The original text is first analysed and disambiguated morphologically (and in the case of deep transfer, syntactically) in order to obtain the source language intermediate representation. The transfer process then converts this final representation (still in the source language) to a representation of the same level of abstraction in the target language. From the target language representation, the target language is generated. [4]

IMAGE 1: ANALYSIS OF A SENTENCE



A. Examples by language

Example 1. I was coming through the eight beautiful gardens.

TABLE 1. ANALYSIS

The sentence	Analysis
I	^Мен<prn><pers><p1><sg><nom>$
eight	^сегіз<num>$
beautiful	^әдемі<adj>$
gardens	^бақша<n><nom>$
through	^арқылы<post>$
was coming	^кел<v><iv><prc_perf>$
	^отыр<vaux><ifi><p1><sg>$

The given model is used to define the parts of speech, hence, for the further specification of the given words position in translation into Kazakh language.

B. Model of the sentence translation

Discrepancy of Kazakh and English language systems generates known problems in translation Kazakh sentences into English: *Мен хат жазып болдым; Мен Астанада болдым* etc. Translating such sentences, it is necessary to specify what is meant: action as the fact, action as action, or – result of some action: Maine Astanada болдым – I was in Astana (last Monday) (exact indication of the period of stay); Мен Астанада болдым – I have been to Astana.

The "chunk" rules are used to translate sentences. The rules are described in the dictionary apertium-eng-kaz.eng-kaz.t1x.

For a simple sentence "I am playing in the garden", the parts of speech are defined first. Then phrases are defined: noun +verb, preposition + noun, etc. The sheaf description "person-verb":

```
<chunk name="pers-verb">
   <tags>
     <tag><lit-tag v="SV"/></tag>
   </tags>
     <lu>
       <clip pos="1" side="tl" part="lem"/>
             <clip pos="1" side="tl" part="a_kaz_verb"/>
       <lit-tag v="aor"/>
       <var n="verb-pers"/>
       <var n="formality"/>
       <var n="verb-nbr"/>
     </lu>
   </chunk>
```

TABLE 2. NOTATIONS

Pos	formality	Verb-pers	verb-nbr
The word position in the sentence is defined	For such verbs in Kazakh language it is required additional suffixes, like „-сіз/сыз"	The verb refers to an animate object, particularly to the person	The auxiliary verb is used

TABLE 1. LEXICAL GROUPS

Meaning Comp	Lexical group
Event	I am playing
Destination	in the garden

I am playing in the garden= Мен бақшада ойнап отырмын.

III. EXPERIMENTS

Here are examples of translations of different sentences in different tenses: Present Simple, Present Continuous, Past Simple, Past Continuous, Future tense.

Present Simple. The sentence: You go to school.

TABLE 4. SENTENCE ANALYSIS

You	**Go to**	**school**
^Сіз<prn><pers><p2><sg><frm><nom>	^бар<v><iv><aor><p2><frm><sg>	^мектеп<n><dat>

TABLE 5. BY SEMANTICS

Designations	Semantics	In dictionary
<prn>	Pronoun	<sdef n="prn" c="Местоимение"/>
<pers>	person	<sdef n="pers" c="Person"/>
<p2>	Which person pronoun	<sdef n="p2" c="2-е лицо"/>
<sg>	number	<sdef n="sg" c="Singular"/>
<frm>	Formal	<sdef n="frm" c="formality 2nd person"/>
<nom>	nominative	<sdef n="nom" c="Nominative"/>

Using various ways of translation [5].

TABLE 6. VARIOUS WAYS OF TRANSLATION

commands	**Result**
Eng-kaz-biltrans	^I<prn><subj><p1><mf><sg>/Мен<prn><pers><subj><p1><mf><sg>$ ^be<vbser><pres><p1><sg>/бол<v><iv><pres><p1><sg>$ ^play<vblex><ger>/ойна<v><tv><ger>$ ^in<pr>/$ ^the<det><def><sp>/<sp>$ ^garden<n><sg>/бақша<n><sg>$ ^.<sent>/.<sent>$
Eng-kaz-debug	Мен бақшада ойнап отырмын
eng-kaz-disam	"<I>" "I" prn subj p1 mf sg SELECT:84; "I" num m sg SELECT:84 "<am>" "be" vbser pres p1 sg "<playing>" "play" vblex ger SELECT:100 ; "play" vblex pprs SELECT:100 ;"play" vblex subs SELECT:100 "<in>" "in" pr "<the>" "the" det def sp "<garden>" "garden" n sg "<.>" "." sent
eng-kaz-interchunk	^subj-pron<SN>{^Мен<prn><pe><p1><sg><nom>$}$ ^prep-det-nom<AdvP>{^бақша<n><lc>$}$ ^pers-verb<SV>{^ойна<v><tv><rc_perf>$ ^отыр<vaux><pres><p1><g>$}$ ^sent<SENT>{^.<sent>$}$
eng-kaz-lex	^I<prn><subj><p1><mf><sg>/Мен<prn><pers><subj><p1><mf><sg>$ ^be<vbser><pres><p1><sg>/бол<v><iv><pres><p1><sg>$ ^play<vblex><ger>/ойна<v><tv><ger>$ ^in<pr>/$ ^the<det><def><sp>/<sp>$ ^garden<n><sg>/бақша<n><sg>$^.<sent>/.<sent>$
eng-kaz-morph	^I/I<num><mf><sg>/I<prn><subj><p1><mf><sg>$ ^am/be<vbser><pres><p1><sg>$ ^playing/play<vb

	lex><ger>/play<vblex><pprs>/play<vblex><subs>$ ^in/in<pr>$ ^the/the<det><def><sp>$ ^garden/garden<n><sg>$ ^./.<sent>$
eng-kaz-postchunk	^Мен<prn><pers><p1><sg><nom>$ ^бақша<n><loc>$ ^ойна<v><tv><prc_perf>$ ^отыр<vaux><pres><p1><sg>$^.<sent>$
eng-kaz-tagger	^I<prn><subj><p1><mf><sg>$ ^be<vbser><pres><p1><sg>$ ^play<vblex><ger>$ ^in<pr>$ ^the<det><def><sp>$ ^garden<n><sg>$^.<sent>$
eng-kaz-transfer	^subj-pron<SN>{^Мен<prn><pers><p1><sg><nom>$}$ ^pers-verb<SV>{^ойна<v><tv><prc_perf>$ ^отыр<vaux><pres><p1><sg>$}$ ^prep-det-nom<AdvP>{^бақша<n><loc>$}$ ^sent<SENT>{^.<sent>$}$

A. *Evaluation and results*

The result of this work is the translator that translates simple sentences in various times - Past Simple, Past Continuous, Present Simple, Present Continuous, Future tense. Comparing work of translators we have come to this results:

B. *Preliminary comparison*

TABLE 7. COMPARISON OF TRANSLATORS

Original	**Apertium**	**Sanasoft**	**Trident**
your two good cities	сіздің екі жақсы қалаңыз	Сенің екі жақсы қалаларың	сендер екі жақсы қалалар
you go to school	Сіз мектепке барасыз	Сіз мектепке бардың	сендер үйрету барасындар.

As you can see from the comparison, Apertium translates precisely 84 %, and Sanasoft 80 % whereas online translator Trident translates correctly only 30 % of the sentence.

Despite the large number of difficulties in translation that were discussed and omitted in this article, it is necessary to understand that such problems arise practically in any language pairs translation[6]. Meaning of system divergences between native and studied languages allow us to predict an interference, to foresee the difficulties and identify weaknesses. Knowing the grammar and the theory of the translation is not suffice to develop skill of correct understanding of the text. Experience shows that to master the translation techniques allocation of certain grammatical and lexical difficulties is necessary and training of their translation is also necessary[6]. Observing languages of the world from the point of grammar, we can see that the logic rules, that regulates process of human thinking, are identical to all languages, and grammar rules for each language are different.

IV. FUTURE WORKS

This paper has described the initial steps to build the prototype of a free/open-source rule-based English–Kazakh machine translation system based on the Apertium platform (apertium-eng-kaz). The current prototype already successfully solves many cases of noun-phrase and adpositional-phrase translation (actually better than the available commercial systems), and contains a reasonable vocabulary for testing purposes, which nevertheless has been completed.

In addition to the immediate actions listed in section 3.2, here is a longer-range set of tasks to be performed in order to have a working machine translation system:

- Completing the coverage of structural transfer rules and monolingual and bilingual vocabularies so that the system produces a translation for at least 90% of the English words and performs the basic operations to identify and process correctly short constituents (1–6 words).
- Releasing the resulting stable system as apertium-eng-kaz version 0.1 and disseminate it to the interested parties to obtain feedback about its functioning. We can reasonably expect this system to work better than the existing commercial systems in most aspects.
- Inserting the resulting system into computer-assisted translation (CAT) workflows being developed at the Universitat d'Alacant: its use to provide edit hints in translation-memory-

based CAT[1] (Esplà-Gomis et al. 2011), and its integration in an Apertium-based interactive machine translation system being developed by Juan Antonio Pérez-Ortiz at the Universitat d'Alacant (this system would provide predictive Kazakh text completions based on the English source text as the professional translators type the translation of that text).

As a longer-range objective, and when a reasonably complete prototype is available, we will tackle another interesting goal: the use of feedback from professional post-editing to improve the system.

Acknowledgment

Acknowledges we thank Francis M. Tyers, Jonathan North Washington and Ilnar Selimcan for their help during the building of this prototype. The help of KazNU master students Aizhan Aitkulova and Anel Jexekova during the first week of work is also gratefully acknowledged.

References

[1] Печерских Т. Ф. Особенности перевода разносистемных языков (на примере английского и казахского языков) [Текст] / Т. Ф. Печерских, Г. А. Амангельдина // Молодой ученый. — 2012. — №3. — С. 259-261.

[2] Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez and Francis M. Tyers: Apertium: a free/open-source platform for rule-based machine translation. In Machine Translation: Volume 25, Issue 2 (2011), p. 127-144.

[3] Sarah Ebling, Andy Way, Martin Volk, & Sundip Kumar Naskar: Combining semantic and syntactic generalisation in example-based machine translation. EAMT 2011: proceedings of the 15th conference of the European Association for Machine Translation, 30-31 May 2011, Leuven, Belgium; eds. Mikel L.Forcada, Heidi Depraetere, Vincent Vandeghinste; pp.209-216 (cites Apertium as a source for "marker words")

[4] Bangalore, S., V. Murdock, and G. Riccardi. 2002.Bootstrapping bilingual data using consensus translation for a multilingual instant messaging system.In Proceedings of 19th International Conference on Computational Linguistics, pages 1–7, Taipei, Taiwan.

[5] Mikel L.Forcada: Free/open-source machine translation: the Apertium platform. Translingual Europe 2010, Hotel Maritim, Berlin, Germany, Monday June 7th 2010; 17pp

[6] Forcada, Mikel L. (2006) "Open-source machine translation: an opportunity for minor languages" in *Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages)* (organised in conjunction with LREC 2006 (22-28.05.2006))

[1] http://www.dlsi.ua.es/~mespla/edithints.html