3rd International Conference on Computer Processing
in Turkic Languages (TURKLANG 2015)

A free/open-source machine translation system for 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

---

Abstract

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-

1 * Corresponding author. *E-mail address:* sun27aida@gmail.com

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.

Keywords: machine translation; Apertium; dictionary; constraint grammar; lexical selection; structural transfer rules

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).

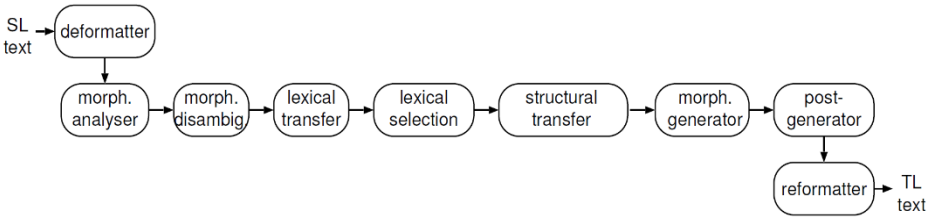

Fig. 1. The pipeline architecture of the Apertium system.

- **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[2] 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.

1.1.2. 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

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

SL word *w*	TL word	Context	Example
			(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*[3], in which will case the phrase will be marked GenP; locative -*{D}{A};* [4] ablative -{*D}{A}н,* etc.; using postpositional constructs based on positional nouns such as *аст* ('under'), *үст* ('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 билеу**ді** *ұнатамын,* 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>

3 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.

4 *{D}* can be *{∂}* or *{m}*, and *{A}* can be *{e}* or *{a}*, depending on the phonological context.

- 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).

1.1.3.2. 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[5], SETimes[6], NewsCommentary[7], Wikipedia[8].

5 *http://www.statmt.org/europarl/v7/es-en.tgz*

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

7 *http://www.statmt.org/wmt13/training-parallel-nc-v8.tgz*

8 *http://dumps.wikimedia.org/enwiki/20140402/enwiki-20140402-pages-articles.xml.bz2*

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%

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[9]	21%	74.52%

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

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
	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.[10]

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.

10 *https://svn.code.sf.net/p/apertium/svn/incubator/apertium-eng-kaz*

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