Әдебиеттер

1. *Жұбанов А.* Автоматты (машиналық аударма) // Аударматану. –Алматы: «Тіл» оқу-әдістемелік орталығы, 2008. –70-93-беттер.

2. *Құрманбайұлы Ш.* Қазақша-орысша, орысша-қазақша терминдер сөздігі (бекітілген терминдер). –Алматы: «Сөздік-Словарь», 2004.

3. *Шарипбаев А.А., Тренкеншу В.П.* Көпсалалы қазақша-орысша-қазақша сөздік. –Астана, 2004.

4. Русско-казахский словарь для государственных служащих. –Астана: «Алтынсофт Астана», 2008.

5. *Қапалбеков Б.С., Құсбекова Б.Ғ., Байменшин А.М., Әбділдаева М.Б.* Орысша-қазақша іскағаз үлгілерінің электронды бағдарламасы. – Алматы: Мемлекеттік тілді дамыту институты, 2010.

6. *Philipp Koehn*. Statical mashine translation. Cambridge University Press, 2009.

7. www.baseage.com

A. SUNDETOVA[1], M.L.FORCADA[2], A. SHORMAKOVA[1], A. AITKULOVA[1].

[1] Information Systems Chair, Al-Farabi Kazakh National University, Al-Farabi av., 71, 050040 Almaty, Kazakhstan, and

[2]Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, E-03071 Alacant, Spain

STRUCTURAL TRANSFER RULES FOR ENGLISH-TO-KAZAKH MACHINE TRANSLATION IN THE FREE/OPEN-SOURCE PLATFORM APERTIUM

Introduction

Translating natural text from English to a Turkic language such as Kazakh faces important challenges:

On the one hand, the complex agglutinative morphology of Turkic languages is very different from that of a fusional, morphologically not too complex language like English; an immediate effect is the fact that correspondences can seldom be modelled as word-for-word translations. Even if Turkic language morphology shows clear morphotactics (ordering of morphemes), its morphophonology shows complex phonological changes to due to interactions between neighboring morphemes (vowel harmony, sonorization, etc.) many of which are explicitly represented in writing.

On the other hand, there are many differences between the syntax of Turkic languages and English. Just to name a few: subject–object–verb order (compare subject–verb–object in English), use of postpositions (compare prepositions in 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), lack of definite articles (extensively used 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, see Печерских & Амангельдина (2012).

When sufficiently large sentence-aligned parallel corpora are available (for instance, as in the case of English to Turkish, see, for example, Tyers and Alperen 2010), statistical machine translation (Koehn 2010) may be used to attempt translation from English into a Turkic language

(in fact, statistical machine translation is currently offered by Google for two Turkic languages, Azeri and Turkish). However, in the case of Kazakh, it would be very hard to put together the necessary amount of sentence-aligned parallel text, and rule-based machine translation, in which experts write up dictionaries and grammatical rules that are applied by an engine, emerges as a clear solution; in fact, existing commercial systems for English to Kazakh (Sanasoft[7], Trident[8]) all appear to be rule-based.

We are currently engaged in building a free/open-source rule-based machine translation system from English to Kazakh, and we are using the Apertium free/open-source machine translation platform (Forcada et al. 2011, http://www.apertium.org) for various reasons. On the one hand, the platform already contains free/open-source English morphological dictionaries and, what is more important, Kazakh morphological dictionaries (Salimzyanov et al. 2013) which take care of all of the morphotactics and morphophonology and provide a basic vocabulary; this allows us to concentrate our work in two fronts: building the lexical transfer part, that is, a bilingual dictionary (already underway) and building structural transfer rules (grammatical rules for translation), which will be the subject of this paper. On the other hand, building free/open-source dictionaries and rules for English to Kazakh means that they will be freely available,[9] for instance, to build translation systems for other Turkic languages; this gives a strategic value to our work, as most of the structural transfer rules will be ready for use with other Turkic languages with little modification or no modification at all.[10]

The paper, which describes work in progress in the Apertium English-to-Kazakh structural transfer, is organized as follows: Section 323 describes the free/open-source rule-based machine translation platform, focusing on structural transfer. Section 0 describes the structural transfer rules currently available to tackle the main syntactic divergences between English and Kazakh; section 0 describes some successful structural translations and some limitations, and, finally, section 0 gives concluding remarks and outlines future work.

The Apertium platform

Apertium (Forcada et al. 2011, http://www.apertium.org) is a free/open-source rule-based machine translation (MT) platform that was launched in 2005 by the Universitat d'Alacant. Though it was initially aimed at translating between closely related languages, it was later extended to be able to deal with unrelated languages. All of the components of the platform (MT engine, developer's tools, and linguistic data for an increasing number of language pairs) are licensed under the free/open-source GNU General Public License (GPL, versions 2 and 3) and are available to everyone interested in the website.

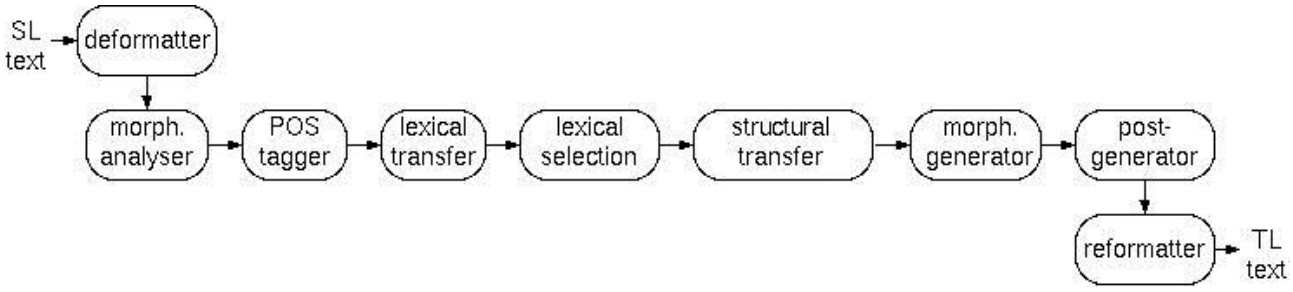

Figure 1: A sketch of the Apertium workflow

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

8 http://www.translate.ua/us/on-line; also through http://itranslate4.eu/en/

9 They already are: see a snapshot at: https://svn.code.sf.net/p/apertium/svn/incubator/apertium-eng-kaz/

10 The Apertium project has a particularly active sub-project for Turkic languages (http://wiki.apertium.org/wiki/Turkic_languages), which has its own mailing list, https://lists.sourceforge.net/lists/listinfo/apertium-stuff.

Apertium-based MT systems are transfer systems implemented as text pipelines (see Figure 1) consisting of the following modules:

1. A deformatter that separates the text to be translated from the formatting tags. Formatting tags are encapsulated as "superblanks" that are placed between words in such a way that the remaining modules see them as regular blanks (for instance, tags in the HTML text I see the sky are encapsulated as I see []the sky[] and everything in square brackets is treated just as regular blanks).
2. A morphological analyser, yielding, for each surface form (SF), for each lexical unit as it appears in the text, a lexical form (LF) composed of: lemma (dictionary or citation form), lexical category (or "part-of-speech"), and inflection information. For instance, the English SF *books* would yield two LFs: *book*, noun, plural, as in *I have bought some books*) or *book*, verb, present tense, 3rd person, as in *He books a ticket*). The morphological analyser executes a finite-state transducer generated by compiling a morphological dictionary for the source language (SL).
3. A constraint-grammar (Karlsson 2005) module based on CG3[11] is used to discard some LFs using simple rules based on context (this module is not depicted in the figure).
4. A part-of-speech tagger based on hidden Markov models (Cutting et al. 1992) selects one of the remaining LFs. The statistical models may be supervisedly trained on an annotated SL monolingual text corpus, or trained in an unsupervised way, either on an unannotated monolingual SL corpus or using two unrelated, unannotated source language and target language corpora (as in Sánchez-Martínez et al. 2008). The Apertium part-of-speech tagger can also read linguistically-motivated constraints (much more rudimentary than constraint grammar rules in the previous module) that forbid specific sequences of two LFs.
5. A lexical transfer module adds, to each source language LF (SL LF), one or more corresponding target language LFs (TL LFs). This module executes a finite-state transducer generated by compiling a bilingual SL–TL dictionary.
6. An (optional) lexical selection module (currently not active in the English→Kazakh system) reads in rules that allow for the selection of one of the TL LFs according to context. When this module is absent, the TL LF given as default in the dictionaries is used.
7. A structural transfer module processes the stream of SL LF–TL LF pairs produced by the lexical transfer module and transforms it into a new sequence of TL LFs; a more detailed description is found in section 0 as this is the main subject of this paper.
8. A morphological generator takes the sequence of TL LFs and generates a corresponding sequence of TL SFs. The morphological generator executes a finite-state transducer generated by compiling a morphological dictionary for the TL.
9. A post-generator takes care of some minor orthographical operations such as apostrophations and contractions in the target language (this module is not used for English to Kazakh).
10. Finally, the deformatter opens the square-bracketed superblanks and places the formatting tags back into the text so that its format is preserved.

Structural transfer in Apertium

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 new 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. Structural transfer rules have a pattern–action form: when a specific (finite-length) pattern of SL LFs is detected, an action builds and generates the corresponding sequence of TL LFs. Rules are applied in a greedy, left-to-right, longest-match fashion. There are two main modalities of structural transfer. The first one (used for related languages) generates the TL LF sequence in a single step. The second one (used in the English–Kazakh system described in this paper) uses three stages to improve the granularity of structural transfer rules (each one has its own rules file):

11 http://beta.visl.sdu.dk/cg3.html

- A first round of transformations ("chunker") detects SL LF patterns and generates the corresponding sequences of TL LFs grouped in chunks representing simple constituents such as noun phrases, prepositional phrases, etc. These chunks bear tags that may be used for inter-chunk processing.
- 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, inter-chunk agreement, 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.

English-to-Kazakh structural transfer

This section describes the current structural transfer in Apertium-eng-kaz (revision 46018, 26.07.2013). English to Kazakh chunker rules (file apertium-eng-kaz.eng-kaz.t1x) are described in detail in section 0. English-to-Kazakh inter-chunk rules (file apertium-eng-kaz.eng-kaz.t2x) are described in detail in section 0. The English-to-Kazakh system has an additional clean-up stage that takes care of the fact that Kazakh morphotactics, as defined in the Kazakh morphological dictionary, contains optional morphemes: for instance, there is no singular morpheme or no-possessive morpheme, but these are generated in the previous three steps. They are eliminated here. Rules for cleanup have the same form as transfer rules for a related language pair (file apertium-eng-kaz.eng-kaz.t4x).

The English-to-Kazakh chunker

As regards the first round of structural transfer (the "chunker", rules written in the apertium-eng-kaz.eng-kaz.t1x file), rules have been written to address some of the the main local morphosyntactic divergences between the languages involved. The prototype is able to perform the local operations necessary to adequately process 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 60, identify six kinds of chunks and translate them into Kazakh as much as possible, leaving some minor operations to be performed in later stages of structural transfer (for instance, the case of noun phrases). Table Table shows a description of the kinds of chunks found, with examples, and the chunk-level tags associated to them. The translation of English noun phrases and prepositional phrases are given below as examples.

Table 2: Chunk-level tags currently associated to each type of chunk. Those marked with an asterisk correspond to optional morphemes that will need special treatment in a cleanup module (see text in section 0).

Phrase	**Description**	**Examples of English chunks detected**	**Chunk-level tags**
NP	Noun phrase	noun determiner–noun numeral–noun adjective–noun determiner–adjective–noun	number*, person, possessor*, case
VP	Verb phrase	finite_verb, do–not–finite_verb have–verb_participle be–verb_gerund must–verb have_to–verb want_to–verb	number, person, tense/conditionality, possessive*, negation*

Phrase	Description	Examples of English chunks detected	Chunk-level tags
PP	Postpositional phrase (except genitive phrases ending in "-{N}{I}ң")[12]	preposition–noun preposition–determiner–noun preposition–numeral–noun preposition–adjective–noun preposition–determiner–adjective–noun	number, person, possessor*, and case
GenP	Genitive postpositional phrase ending in "-{N}{I}ң"	(same as PP, with "of" as preposition)	number, person, possessor*, and case
AdjP	Adjectival phrase (except superlatives with "ең")	adjective "more"–adjective	(none)
SupP	Superlative phrase (adjectival phrase with "ең")	adjective_"-est" the–"most"–adjective	possessor*,[13] case

Translation of noun phrases

Consider the following example: the chunker identifies the English sequence *the large book* (determiner–adjective–noun) as a noun-phrase chunk. It translates into Kazakh, and assigns it four chunk-level tags: number (set to singular), person (set to 3rd), possessor (to be determined, as the noun *кітап* ('book') could receive a 3rd-person possesive ending (кітабы) later if the context were, for instance, *the large book of animals, аңдардың үлкен кітаб**ы***), and case (to be determined as it could be, for instance, accusative in *I saw the large book, Мен үлкен кітап**ты** көрдім*).

Translation of prepositional phrases

On encountering an English prepositional phrase, which has to be rendered in Kazakh as a postpositional phrase, there are three possible outcomes:

(1) The prepositional phrase results in a simple postpositional phrase using the locative "-{D}{A}",[14] ablative "-{D}{A}н", etc., but not the genitive "-{N}{I}ң":

[PP [P *in*] [NP *the beautiful garden*]] → [PP [NP *әдемі бақша*] [P -*да*]]

(2) The prepositional phrase results in a simple postpositional phrase using the genitive -NIң, which will be marked GenP:

[PP [P *of*] [NP *the beautiful garden*]] → [GenP [NP *әдемі бақша*] [P -*ның*]]

(3) The prepositional phrase results in a complex postpositional phrase based around a noun such as *аст*, *үст*, etc.:

[PP [P *under*] [NP *the garden*]] → [PP [NP [GenP [NP *бақша*][P -*ның*]] [NP *астын*]] [P -*да*]]

In all three cases, the possessor tag of the chunk, which corresponds to the main noun in the PP or the GenP has to be left open to being determined in later transfer operations (consider, for instance, the case that the PP *in the beautiful garden* is part of a larger structure, *in the beautiful garden of the city*, *қаланың әдемі бақша**сын**да*, in which the noun *бақша* 'garden' receives a possessive ending).

Translation of verb phrases

The mapping of English verb tenses onto Kazakh is not completely straightforward and is treated in the chunker. Just to give a few examples, present simple and future are rendered using the same

12 Upper case letters in braces (such as {N}) represent hypothetical archiphonemes (actually archigraphemes) that are realized as phonemes (actually graphemes) after morphophonological rules have been applied. For instance, in the genitive ending "-{N}{I}ң", the archiphoneme {N} may be realized as т, д, or н and the archiphoneme {I} may be realized as і or ы depending on the previous phonological context. This is perfromed during morphological generation (see section 0).

13 Treated as a noun phrase with an implied noun ("the largest [book]")

14 {D} can be *д* or *т*, and {A} can be *е* or *а*, depending on the phonological context.

tense in Kazakh (*I play → Мен ойнаймын; I will play → Мен ойнаймын*); 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*) where the main verb is found in the *-п* participle form (*ойнап*), and a suitable finite form (*жатырмын, отырдым*) of an auxiliary verb (*жатыр, отыр*) is used to express number and person agreement[15] (see Table Table) for details.

Table 3 Examples of tense mapping operations performed at the chunk level

English tense	Example	Morphemes	Equivalent tense in Kazakh	Translation
Present Simple	*I play*	*ойна* + *й*(*а* or *е*) <aorist> + <person>	Ауыспалы осы шақ (changing present simple)	*Мен ойнаймын*
Present Continuous	*I am playing*	*ойна* + *п* (*ып* or *ін*) <perfect participle> + *жатыр* + <present> + <person>	Нақ осы шақ (now present tense)	*Мен ойнап жатырмын*
Past Continuous	*I was playing*	*ойна* + *п* (*ып* or *ін*) <perfect participle> + *отыр* + *д{I}* <past> + <person>	Бұрынғы өткен шақ (past continuous)	*Мен ойнап отырдым*

Verb-phrase chunks are also used to prepare translations not using a finite verb (but a nominal or adjectival structure, often based on non-finite forms of verbs instead). For instance, for obligatory English modal constructs (*have to, must, need to, should*) verb phrases made up of three lexical units have to be generated, with a verbal noun, an adjective roughly meaning "necessary" (*керек*) or "proper" (*жөн*), and a form of the copula (absent in present tense); the subject receives the genitive or dative case: *I have to go → Менің баруым керек*, *I need to go → Маған бару керек*, etc.; see these and other modal construction examples in Table Table.

Table 4 Translation of some English modal verbs

Construction	Example	Morphemes	Translation	Gloss
Must have to	*I must go, I have to go*	*Мен* + *-{N}{I}ң* <genitive> + *бар* + *-у* <gerund> + *-{I}м* <1st person possessive> + *керек* <adjective>	*Менің баруым керек*	My going necessary [is]
Should	*I should go*	*Мен* + *-{N}{I}ң* <genitive> + *бар* + *-{G}{A}н* <past gerund> + *-{I}м* <1st person possessive> + жөн <adjective>	*Менің барғаным жөн*	My going proper [is]
Need to	*I need to go*	*Мен* + *-{G}{A}[н]* <dative> + *бар* + *-у* <gerund> + *керек* <adjective>	*Маған бару керек*	To me, going necessary [is]
Want to	*I want to go*	*Мен* + *-{N}{I}ң* <genitive> + *бар* + *-{G}{I}* + *-{I}м* <1st person possessive> + *кел* + *-{E}д{I}* <past, 3rd person>	*Менің барғым келеді*	My going will come

15 Actually, Kazakh language uses four auxiliary verbs: *жатыр* ('lie', used when the activity takes a long time), *отыр* ('sit', used when the activity appears to be done in a sitting position), *тұр* ('stand', when the takes a short time), and *жүр* (when the activity repeats regularly). Choosing the most adequate auxiliary verb is hard without a semantic analysis, which is not easily available in Apertium. Our current choice (an approximation) is *жатыр* ('lie') for the present continuous and *отыр* ('sit') for the past continuous.

Finally, as negative constructions in English contain more words than their corresponding affirmative words, or may even use an auxiliary verb (as in *do not*, *did not*), they have to be separately detected as verb chunks to generate the appropriate Kazakh negative forms (*I play → мен ойнаймын; I do not play → мен ойна***ма***ймын*. For examples of other negative constructs, see Table Table).

Table 5 Translation of some negative constructions.

Construction	Example	Morphemes	Translation	Note
Present Continuous (negative)	*I am not playing*	*ойна* + *п* (*ып* or *ип*) <perfect participle> + *жатқан жоқ* + <person>	*Мен ойнап жатқан жоқпын*	In present auxiliary verbs (жатыр/отыр) do not have a synthetic negative form.
Can (negative)	*I can not play*	*Ойна* + *-{E}* <imperfect participle> + *ал* + *ма* <negative> + *й* <aorist> + *мын* <1st person>	*Мен ойнай алмаймын*	

Verb phrases (VP) are marked at the chunk level with person and number, both to be determined and linked via references to the appropriate morphemes in the appropriate verb lexical forms. The chunk-level person and number to be determined will be rewritten by the appropriate 2nd-level (interchunk) transfer rules, and will be propagated to lexical forms at the 3rd-level transfer stage (postchunk).

Other indicators that have to be made available at the chunk level are negation (for negative verbs) and conditional (which will be handled as a tense). For instance, negation can be easily determined at the chunking level when the English VP chunk contains *not*, as in *I don't play → мен ойна***ма***ймын*, but may need to be determined at the interchunk level in sentences having a non-negative VP but a negative word like those starting with *еш-*, like *I write nothing → мен ешнәрсе жаз***ба***ймын*, which requires a negative form of the verb (*-ба-* in the example).

Translation of adjectival phrases

In Kazakh noun phrases, adjectives come before nouns and do not show any agreement with nouns. Adjectives can also appear in separate adjective phrases. Here are some examples:

(4) The adjective alone, marked AdjP:

[AdjP *beautiful*] → [AdjP *әдемі*]

(5) Comparative adjective phrases (English *more* + adjective, or adjective-*[e]r*); the Kazakh translation chooses the comparative suffix "-{I}p{A}{K}":

[AdjP *more beautiful]* → [AdjP *әдем***ірек**]

(6) For superlative adjective phrases "*the most* + adjective" or "adjective-[*e*]*st*", translation is built using "ең" + adjective:

[SupP *the largest]* → [SupP **ең** *әдемі*]

[SupP *the most beautiful]* → [SupP **ең** *үлкен*]

As noted in §0, superlative adjective phrases have some properties of noun phrases (such as receiving possessive morphemes when modified by a genitive phrase: *the most beautiful of people → адамдардың ең әдемі***сі**); one could say that they are treated as NPs with an implied noun.

English-to-Kazakh inter-chunk processing

The second round of structural transfer (the "interchunk" rules written in the apertium-eng-kaz.eng-kaz.t2x file) is currently performed by a proof-of-concept set of 18 rules, representative of following operations:

- Inter-chunk agreement (for instance, number and person agreement between subject noun phrase and verb phrase): features to be agreed here are left undefined by the chunker; those that are not defined at the interchunk phase are left for the post-chunk phases.

- Assigning case to noun phrases (which are generated without case by the chunker): for instance, accusative case for objects (*I see the sky → Мен аспан***ды** *көремін*), genitive case for obligatory constructs (*I have to go → мен***ің** *баруым керек*), dative case for the verb *to need* (*I need a book →Ма***ған** *кітап керек*), locative case for possession (*I have a book → Мен***де** *кітап бар*), etc.
- Reordering: placing of object before verb (*I[1] see[2] the sky[3] → Мен[1] аспанды[3] көремін[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.

The set of rules has to be extended, as many combinations of the above phenomena are still not covered (for instance, there is no rule to obtain the right word order in *I have to go to the university →Менің университетке баруым керек*).[16]

Some results, problems and limitations

The system described is not much more than a proof-of-concept system that still needs to be extended to reasonably cover all transfer operations needed. Therefore, evaluating the output of the system using customary evaluation measures such as BLEU (Papineni et al. 2002) is still out of the question.

Instead, tables Table and Table show how our current prototype performs for some representative structures covered by the transfer rules currently available (some of them discussed above). As has been said above, are already at least two MT systems that translate from English to Kazakh: Sanasoft's and Trident's, both of which can be used online (see Introduction for details); therefore, we will briefly compare our results to those obtained by the commercial systems.

Table 6 Example machine translation output for some simple phrases and sentences.

Structure/problems	**English**	**Kazakh (Apertium)**	**Kazakh (Sanasoft)**	**Kazakh (Trident)**
Noun phrases	*your two beautiful gardens*	*сіздің екі әдемі бақшаңыз*	*Сенің екі әдемі бақ***тарың**.	**сендер** екі тамаша бақша**лар**
Prepositional phrases	*in the big city*	*үлкен қалада*	*Үлкен* **қала**	*үлкен қалада*
Possessives	the chief of the city	*қаланың басшысы*	*қаланың көсемі*	**Бас қала**
	On top of the tree of the garden of the city	*қала бақшасының ағашының үстінде*	**Зырылдауық ағаш бақ қала**	**В алқындыр-қаланың бақшасының ағашының**
Adjective phrases	bigger	*үлкенірек*	*Үлкен*	**үлкен**
Modal verbs	I have to go	*Менің баруым керек*	**Мен барып жатырмын жүрмін**	*Маған бару* **have**
	I can drive	*Мен жүргізе аламын*	*Мен* **болып жатырмын** *жүргізіп* **жатырмын**	*Мен жүру* **білемін**

16 As chunks detected by the chunker are finite-length and inter-chunk rules also process finite-length chunk sequences, it has to be noted that there will always be a limit to the scope of reordering or agreement rules.

Concluding remarks and future work

The current prototype already successfully solves many cases of noun-phrase, verb-phrase, prepositional-phrase, and adjectival-phrase translation (some actually better than the available commercial systems), and contains a reasonable vocabulary for testing purposes, which nevertheless still needs extending for real-world applications.

The following tasks have 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 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.

As a longer-range objective, and when a reasonably complete prototype is available, we will tackle another interesting goal: the use of feedback from human input (for instance, in an interactive machine translation system that provides completions to what the translator is typing).

Table 7: Example machine translation output for some simple phrases and examples.

English	Kazakh (Apertium)	Kazakh (Sanasoft)	Kazakh (Trident)
I see the blue sky.	*Мен көк аспанды көремін*	*Мен**де** көк аспан көріп жатырмын*	*Мен көгілдір аспанды көремін.*
You go to school	*Сіз мектепке барасыз*	*Сіз мектепке бар**дың***	*сендер үйрету барасыңдар*
A book has been given to you	*кітап сізге беріліп болған*	*Кітап* **барып жатыр** сізге берсін	Кітап **жібер-** сендерге болды
I can go to the three big shop	*Мен үш үлкен дүкенге бара аламын*	*Мен* **three** *үлкен магазинге* **болып** *жатырмын жүрмін*	*Мен* **үшке** *деген бару үлкен дүкен білемін*
The most beautiful of garden is opened	*бақшаның ең әдемісі ашылады*	**Көпшілік** *әдемі бақ ашық* **бар**	*Ең тамаша бақш**адан** болады* **ашыл-**
I see my car	*Мен менің жеңіл автокөлігімді көремін*	*Мен менің автомобилім көріп жатырмын*	*Мен өзінің автомобильсын көремін*
The famous doctor of the city is going to hospital	*қаланың танымал дәрігері емханаға барып жатыр*	*Атақты дәрігер қала ауруханаға барады*	*қаланың* **атайы** *докторы ауруханаға деген жиналады*
She eats chocolates with sugar	*Ол шоколадтарды қантпен жейді*	*Ол* **eats chocolates** *қант*	*Ол шоколадтарды қантпен жейді*

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.

References

1. Cutting, D., Kupiec, J., Pedersen, J., Sibun, P. 1992. “A practical part-of-speech tagger”. *Proceedings of the 3rd Conference on Applied Natural Language Processing* (ANLP’92), p. 133-140.

2. 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.

3. Karlsson, F., Voutilainen, A., Heikkilä, J., Anttila, A. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.

4. Koehn, P. (2010) *Statistical Machine Translation*, Cambridge University Press.

5. 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/]

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

7. Salimzyanov, I., Washington, J.N., Tyers, F.M. “A free/open-source Kazakh-Tatar machine translation”. Proceedings of MT Summit XIV (Nice, France, 4–6 September 2013), accepted.

8. Sánchez-Martínez, F., Pérez-Ortiz, J.A., Forcada, M.L. (2008). “Using target-language information to train part-of-speech taggers for machine translation”. *Machine Translation*, 22(1-2)29–66.

9. Tyers, F. M. and Alperen, M. S. (2010) “SETimes: A parallel corpus of Balkan languages”. Proceedings of the MultiLR Workshop at the Language Resources and Evaluation Conference at LREC2010, 49–53 [http://www.lrec-conf.org/proceedings/lrec2010/workshops/W22.pdf].

У. КАМАНУР, Б.З АНДАСОВА, Б.М БАЙГУШЕВА

Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан

ҚАЗАҚ-АҒЫЛШЫН-ҚЫТАЙ ДЫБЫСТЫҚ СӨЗДІГІН ӘЗІРЛЕУ

Қазақ тілі Қазақстан Республикасының мемлекеттік тілі ретінде елдің ішкі қарым-қатынасында ғана емес, сонымен бірге халықаралық байланыстарда да қолданылуы керек. Қазақстан Республикасы дүниежүзілік қауымдастыққа біріге отырып мемлекеттік тілдің даму деңгейін оның қарқынды зерттелуі мен қазақ тілін оқытудың әдістері мен құралдарын жасау негізінде дүниежүзілік деңгейге жеткізуді қамтамасыз ету керек. Сондықтан да қазіргі таңда тілді меңгеруде, көптілді қарым-қатынас дамыған заманда көптілді дыбыстық сөздікті құру технологиясын жасау өзекті болып отыр.

Бүгінгі таңда сөйлеу технологияларын қолдану арқылы дыбыстық сөздіктерді жасау қарқынды түрде дамып келеді. Бірақ олар негізінен ағылшын, орыс, қытай және т.б. тілдерге бағытталған. Қазақ тіліне арналған жарамды сөйлеу технологияларымен жасалған дыбыстық сөздіктер жоқтығы қазақ тілін бұл заманауи жетістіктерден тыс қалдырады.

Осы мақсатта ақпараттық және байланыс технологиялары саласы бойынша қазақ тіліндегі терминдер мен олардың ағылшын, орыс және қытай тілдеріндегі аудармаларының 4000 бірлік көлеміндегі базасы, қазақша сөйлеуді синтездеуге арналған дифон базасы құрастырылып, көп тілді дыбыстық сөздіктің онтологиясы, сонымен қатар, сөздікке енгізу үшін сөздерді тану және дыбыстық сөздіктегі сөздерді синтездеу алгоритмдері мен программалары жасалды.

Қазіргі кезде біздің елімізде бірнеше электрондық сөздіктер бар. Бірақ олардың ішінде әлі күнге дейін жаңа электрондық сөздіктерді жасауға немесе бар электрондық сөздіктерді өңдеуге мүмкіндік беретін компьютерлік программа жоқ. Белгілі бір салада сөздік жасаушылар өзі жасаған сөздіктің электрондық үлгісін алу үшін және оны кеңінен таратуға мүмкіндік беретін тиісті программа жоқ. Сонымен қатар осындай сөздіктердің ақырғы