A free/open-source Kazakh-Tatar machine translation system

Ilnar Salimzyanov
Kazan Federal University
Kazan, Republic of Tatarstan
Russian Federation
ilnar.salimzyan@gmail.com

Jonathan North Washington
Departments of Linguistics
and Central Eurasian Studies
Indiana University
Bloomington, Indiana 47405 USA
jonwashi@indiana.edu

Francis Morton Tyers
Departament de Llenguatges
i Sistemes Informàtics
Universitat d'Alacant
E-03877 Alacant
ftyers@dlsi.ua.es

Abstract

This paper presents a bidirectional machine translation system between Kazakh and Tatar.

1 Introduction

This paper presents a prototype shallow-transfer rule-based machine translation system between Kazakh and Tatar.

The paper will be laid out as follows: Section 3 gives a brief description of the two languages; Section 2 gives a short review of some previous work in the area of Turkic–Turkic language translation; Section 4 describes the system and the tools used to construct it; Section 5 gives a preliminary evaluation of the system; and finally Section 6 describes our aims for future work and some concluding remarks.

2 Previous work

Within the Apertium project, work on MT systems between Turkic languages has been started (Turkish-Kyrgyz, Azeri-Turkish), but the Kazakh-Tatar system described by the present study is the closest to production-ready of them. Among these systems is a prototype Tatar-Bashkir machine translation system which was built by the authors of this paper (Tyers et al., 2012); due to the closeness of these languages, it proved to provide high accuracy in its translations, but being a prototype system by design, had relatively low coverage.

Besides these systems, several previous works on making machine translation systems between Turkic languages exist, although to our knowledge none are publicly available. Some MT systems have been reported that translate between Turkish and other Turkic languages, including Turkish-Crimean Tatar (Altintas, 2001b), Turkish-Azerbaijani (Hamzaoğlu, 1993), Turkish-Tatar (Gilmullin, 2008), and Turkish-Turkmen (Tantuğ et al., 2007), though none of these have been released to a public audience.

3 Languages

Both Tatar and Kazakh belong to the Kypchak (or North-western) group of Turkic languages. The spoken and written languages share some level of mutual intelligibility to naïve native speakers, though this is somewhat limited, and is obscured by different orthographical conventions and some opaque correspondences.

Kazakh is primarily spoken in Kazakhstan, where it is the national language, sharing official status with Russian as an official language. Large groups of native speakers also exist in China, neighbouring Central-Eurasian republics, and Mongolia. The total number of speakers is at least 10 million people.

Tatar is spoken in and around Tatarstan by approximately 6 million people. It is co-official with Russian in Tatarstan — a republic within Russia. A majority of native speakers of both languages are bilingual in Russian.

3.1 Phonological differences

As closely related languages, Kazakh and Tatar share many phonological processes, including front-back vowel harmony systems, consonant voicing assimilation, and even a typologically rare consonantal nasal harmony system. However, the differing details of these processes and the existence of processes unique to each language render Kazakh and Tatar fairly different. For example, Kazakh has a ubiquitous system of desonorisation of the initial sonorants found in many common morphemes. Furthermore, Tatar has nasal assimilation of the initial /l/ of the plural-suffix.

3.2 Orthographic differences

The standard varieties of Kazakh and Tatar our system deals with are both written in Cyrillic, though their implementations of Cyrillic differ in many ways.

While Tatar and Kazakh both have a velar/uvular obstruent distinction (e.g., /k/ vs. /q/) that interacts with adjacent vowels, the Tatar orthography only has one series of letters (e.g., ‹к›), relying on adjacent vowels (and employing "hard" and "soft signs" when these fail) to differentiate the two, and Kazakh has two series of obstruents (e.g., ‹к› and ‹қ›).

Kazakh does not orthographically distinguish high unrounded vowels (/ɘ/ ‹і› and /ə/ ‹ы›) before glides (/w/ ‹у› and /j/ ‹й›) by writing the combination with one letter; i.e., /ɘj/ and /əj/ are both written ‹и›, while /ɘw/ and /əw/ are both written ‹у›. The quality of these vowels is necessary to know in order to predict the quality of following harmonising vowels. Additionally, Tatar and Kazakh both use "yoticised" vowels—i.e., when ‹о›, ‹у›, or ‹а› (along with ‹ә› in Tatar) follow /j/, a single character is used to represent both: ‹ё›, ‹ю›, and ‹я› respectively.[1]

All of these orthographical conventions present accute challenges to designing accurate morphological transducers for the languages.

3.3 Morphological differences

There are a number of examples where the morphologies of Kazakh and Tatar are rather different, including morphemes in one language that don't exist in the other, entirely different uses of the same morpheme combinations, and morphotactic differences (i.e., allowable ordering and placement of morphemes).

An example of a morpheme that doesn't exist in one of the two languages is Kazakh `-{E}т{I}н`, which is used to form non-past verbal adjectives and verbal nouns. The semantically equivalent structure in Tatar is `-{E} торган`, which historically corresponds to the source of the Kazakh morpheme; however, the use of `-{E} тұрған` in modern Kazakh is different from that of `-{E}т{I}н`.

Another example of a far-reaching morphological difference between Tatar and Kazakh is the presence of a four-way distinction in Kazakh's 2nd person system (both pronouns and agreement suffixes), where Tatar only has a two-way distinction. Kazakh has a distinct pronoun for all combinations of [±plural, ±formal], whereas Tatar collapses all pronouns except the [-plural, -formal] into one pronoun, as summarised in table 1.

This systematic difference would seem to be a minor issue, since, as is typical in pro-drop languages, pronouns are only used for emphasis and clarification. However, this difference between Tatar and Kazakh in the second person system runs much deeper than just the pronoun system. Since all finite verb forms morphologically agree in person and number with their subject and all possessed nouns agree in person and number with their possessor (even when there is no overt pronoun, in either situation), the Kazakh and Tatar systems of agreement suffixes reflect the same pattern; i.e., there are several sets of agreement morphemes which have a one-to-one correspondence with the pronouns in each language, resulting in several systems of suffixes in each language that have the same set of distinctions as in the 2nd person pronoun systems.

	-pl	+pl
-formal	сен	сендер
+formal	сіз	сіздер

(a) Kazakh 2nd person pronouns

	-pl	+pl
-formal	син	сез
+formal	сез	сез

(b) Tatar 2nd person pronouns

Table 1: The 2nd person pronoun systems of Kazakh and Tatar

The past tense systems of Kazakh and Tatar have a many-to-many correspondence. As shown in table 2, at a basic level, in the past tense, Kazakh differentiates [±eyewitness][2] (where [-eyewitness] is used for cases of both potentially unreliable information and newly discovered information) and [±recent], whereas Tatar has only three categories: eyewitness, non-eyewitness, and newly-acquired-information—all with no [±recent] distinction. As an example of the many-to-many correspondence that this results in, Tatar has a single non-eyewitness past tense morpheme (`-GAн-`) while Kazakh has a recent non-eyewitness past (`-Iп-`) and a distant non-eyewitness past (`-GAн екен-`). On the other hand, these two non-eyewitness past forms in Kazakh are used for both potentially unreliable information and newly acquired information, whereas in Tatar, non-eyewitness (`-GAн-`) and newly-acquired-information (`-GAн- икән`) past forms are distinguished.

Without regard to the semantic alignment of these forms, the morphotactics of the cognate Kazakh distant non-eywitness past (`-GAн екен-`) and Tatar newly-acquired-information past (`-GAн- икән`) are different. Specifically, in both languages, the person agreement takes the form of a person copula suffix, although in Kazakh this suffix follows the tense morphemes (e.g., барған екенсің "apparently you went"), whereas in Tatar this suffix intervenes between the two pieces of the

[1] Furthermore, in Tatar, /j/ followed by ‹э› or ‹ы› in Tatar is represented by ‹е›, though ‹е› is also the non-word-initial variant of ‹э›.

[2] "Eyewitness" is a convenient term for this feature, though it may be better expressed as simply "reliability of knowledge" (which indeed often equates to whether the knowledge was acquired first-hand or not) in many cases.

	[+recent]	[-recent]
[+reliable]	-DI-	-GAн-
[-reliable]	-Iп-	-GAн екен-

(a) Kazakh past tense morphology

	[-newlyAcq'd]	[+newlyAcq'd]
[+reliable]	-DI-	-GAн- икән
[-reliable]	-GAн-	

(b) Tatar past tense morphology

Table 2: A comparison of the basic past-tense morphology of Kazakh and Tatar

"compound" tense morpheme (e.g., баргансың икән "I guess you went").

Another morphotactic difference between Kazakh and Tatar is found with the negative forms of the cognate `-GAн-` past tenses. In Kazakh, the negative form of the non-recent reliable-information past tense is `-GAн емес-`, whereas in Tatar, the negative form of the non-eyewitness past tense is `-мAGAн-`.

3.4 Syntactic differences

There are a number of syntactic differences in Tatar and Kazakh, which include differences in verb valencies in equivalent translations, FIXME, and FIXME.

An example of a difference in verb valencies is with the expression corresponding to "to like something" in Kazakh and Tatar. In Kazakh, the verb ұна is used, e.g. бауырсақ маған ұнайды "I like bawyrsaq", where the subject "I" in English is expressed through a dative experiencer in Kazakh and the object "bawyrsaq" in English is the grammatical subject in Kazakh. Tatar, on the other hand, uses a verb whose arguments correspond to the arguments of "to like" in English, e.g. мин бавырсак яратам "I like bawyrsaq", where the first person pronoun is the grammatical subject and "bawyrsaq" is the grammatical direct object (with no accusative suffix since it is indefinite).

In Kazakh, a gerund (i.e., verbal noun) with case marking and sometimes person agreement in the form of possessive suffixes is used to make a verb phrase an argument to certain other main phrases. In Tatar, many of these phrases use an invariant infinitival form. Some examples are shown below:

(1) a. Мен үйге қайтуым керек.
мен үй-GA қайт-у-Iм керек
I home-DAT go-GER-1SG need
'I need to go home.'

b. Миңа өйгә кайтырга кирәк.
мин-GA өй-GA кайт-IргA кирәк
I-DAT home-DAT go-INF need
'I need to go home.'

> Tatar infinitive (Kazakh ger+nom/acc/dat, Kazakh -GAlI vadv)

4 System

The system is based on the Apertium machine translation platform (Forcada et al., 2011).[3] The platform was originally aimed at the Romance languages of the Iberian peninsula, but has also been adapted for other, more distantly related, language pairs. The whole platform, both programs and data, are licensed under the Free Software Foundation's General Public Licence[4] (GPL) and all the software and data for the 30 supported language pairs (and the other pairs being worked on) is available for download from the project website.

4.1 Architecture of the system

The Apertium translation engine consists of a Unix-style *pipeline* or *assembly line* with the following modules (see Fig. 1):

- A *deformatter* which encapsulates the format information in the input as *superblanks* that will then be seen as blanks between words by the other modules.
- A *morphological analyser* which segments the text in surface forms (SF) (*words*, or, where detected, multi-word lexical units or MWLUs) and for each, delivers one or more *lexical forms* (LF) consisting of *lemma*, *lexical category* and morphological information.
- A *morphological disambiguator* (constraint grammar) which chooses, using linguistic rules the most adequate sequence of morphological analyses for an ambiguous sentence.
- A *lexical transfer* module which reads each SL LF and delivers the corresponding target-language (TL) LF by looking it up in a bilingual dictionary encoded as an FST compiled from the corresponding XML file. The lexical transfer module may return more than one TL LF for a single SL LF.
- A *lexical selection* module which chooses, based on context rules the most adequate translation of ambiguous source language LFs.
- A *structural transfer* module which performs local syntactic operations, is compiled from XML files

[3] http://www.apertium.org
[4] http://www.fsf.org/licensing/licenses/gpl.html

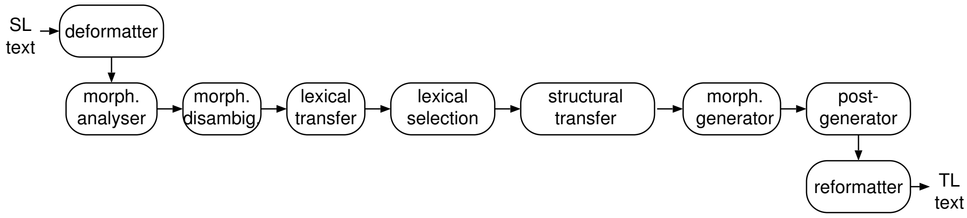

Figure 1: The pipeline architecture of the Apertium system.

containing rules that associate an *action* to each defined LF *pattern*. Patterns are applied left-to-right, and the longest matching pattern is always selected.

- A *morphological generator* which delivers a TL SF for each TL LF, by suitably inflecting it.
- A *reformatter* which de-encapsulates any format information.

4.2 Morphological transducers

The morphological transducers are based on the Helsinki Finite State Toolkit (Linden et al., 2011), a free/open-source reimplementation of the Xerox finite-state toolchain, popular in the field of morphological analysis. It implements both the **lexc** formalism for defining lexicons, and the **twol** and **xfst** formalisms for modeling morphophonological rules. It also supports other finite state transducer formalisms such as **sfst**. This toolkit has been chosen as it — or the equivalent XFST — has been widely used for other Turkic languages (Çağrı Çöltekin, 2010; Altintas, 2001a; Tantuğ et al., 2006; Washington et al., 2012; Tyers et al., 2012), and is available under a free/open-source licence.

The morphologies of both languages are implemented in lexc, and the morphophonologies of both languages are implemented in twol.

Use of lexc allows for straightforward definition of different word classes and subclasses. For example, Tatar (but not Kazakh) has two classes of verbs: one which takes a harmonised high vowel in the infinitive (the default), and one which take a harmonised low vowel in the infinitive. This was implemented in lexc with two similar continuation lexica for verbs: one pointing at a lexicon with an A-initial infinitive ending, and another pointing at a lexicon with an I-initial infinitive ending.

Use of twol allows for phonological processes present in the languages, like vowel harmony and desonorisation, to be implemented in a straightforward manner. For example, in Tatar, the A and I archiphonemes found in the infinitive are harmonised to one of two vowels each, depending on the value of the preceding vowel; the basic form of this process can be implemented in one twol rule.

The same morphological description is used for both analysis and generation. To avoid overgeneration, any alternative forms are marked with one of two marks, LR (only analyser) or RL (only generator). Instead of the usual compile/invert to compile the transducers, we compile twice, once the generator, without the LR paths, and then again the analyser without the RL paths.

4.3 Bilingual lexicon

The bilingual lexicon currently contains 9,269 stem-to-stem correspondences and was built mostly by hand (i.e., by translating Kazakh stems unrecognized by the morphological analyser into Tatar). Some toponyms and other proper names were translated semi-automatically by looking up links in Wikipedia; also, some Russian loanwords common to both languages (such as автомобиль, гонорар, etc.) were added to the bilingual dictionary automatically by taking the intersection of Russian and Kazakh wordlists.

Entries consist largely of one-to-one stem-to-stem correspondences with part of speech, but also include some entries with ambiguous translations (see e.g., Fig. 2).

4.4 Disambiguation rules

The system has a morphological disambiguation module in the form of a Constraint Grammar (CG) (Karlsson et al., 1995). The version of the formalism used is vislcg3.[5]

The grammar currently has 58 rules. The goal of these rules is to select the correct analysis when there are multiple morphological analyses of Kazakh forms.

The output of the morphological analyser is highly ambigious. Both Kazakh and Tatar have a lot of affixes which join the verb stem. Some of the grammars label them all gerunds, or participles, assuming that there are no finite forms.

We give to a VerbStem+ParticularAffix surface form maximal number of readings (syntactic roles) it can

[5] http://beta.visl.sdu.dk/constraint_grammar.html

```
<e><p><l>құрал<s n="n"/></l><r>корал<s n="n"/></r></p></e>
<e><p><l>құрал<s n="n"/></l><r>чара<s n="n"/></r></p></e>
<e><p><l>есім<s n="n"/></l><r>исем<s n="n"/></r></p></e>
<e><p><l>топ<s n="n"/></l><r>туп<s n="n"/></r></p></e>
<e><p><l>топ<s n="n"/></l><r>төркем<s n="n"/></r></p></e>
<e r="RL"><p><l>топ<s n="n"/></l><r>группа<s n="n"/></r></p></e>
```

Figure 2: Example entries from the bilingual transfer lexicon. Kazakh is on the left, and Tatar on the right

have, such as [verbal adjective], [finite form], [substantivized verbal adjective, i.e., if the word it determines was ommitted], whereas many grammars do not distinguish these and label forms of verbs containing a particular affix just as a certain gerund or participle form.

This type of ambiguity is common between the two languages we deal with, and therefore may be passed from one language to another and most likely won't lead to translation errors. Disambiguating it, and, prior to that, having that kind of informaton about the syntactical role of a surface form is crucial for success e.g. of a Turkic to English system.

This is the reason for a high ambiguity rate of the transducers.

4.5 Lexical selection rules

While many lexical items have a similar range of meaning, lexical selection can sometimes be problematic between Kazakh and Tatar.

For example (see figure 2), Kazakh құрал can mean an instrument, device, tool, or even weapon, all meanings corresponding to its Tatar cognate корал; however, it is also used in the compound ақпарат құралдары 'mass media' (literally, 'means of information'), which translates to Tatar as мәгълүмат чаралары (which has the same literal translation). Hence, the Kazakh word құрал must have two entries in the bilingual lexicon: one that corresponds to Tatar корал and one that corresponds to Tatar чара. A lexical selection rule that selects the translation чара when it occurs in a compound with ақпарат is written to ensure the correct translation; this rule is shown in figure 3.

Likewise, the Kazakh word топ has two translations into Tatar. One translation is туп 'ball' (which can also be доп in Kazakh), and the other translation is төркем 'group'. The bilingual dictionary also has the Russian word группа 'group', which is used in Tatar, as an entry which may be translated to Kazakh топ (i.e., analysed), but is never generated.

The system currently has a total of 33 lexical selction rules.

5 Evaluation

Lexical coverage of the system is calculated over freely available corpora of Kazakh and Tatar.

For Kazakh, two years worth of content (2010 and 2012) from Radio Free Europe / Radio Liberty's Kazakh-language service,[6] as well as a recent dump of Wikipedia's articles in Kazakh[7] were used.

For Tatar, the Tatar Wikipedia,[8] the New Testament in Tatar, and content from Radio Free Europe / Radio Liberty's Tatar-language service[9] from early 2007 to early 2012.

The versions of the transducers tested were `r43595` from the Apertium SVN.[10] Corpora were divided into 10 parts each; the coverage numbers given are the averages of the calculated percentages of number of words analysed for each of these parts, and the standard deviation presented is the standard deviation of the coverage on each corpus.

As shown in table 4, the naïve coverage of the Kazakh-Tatar MT system[11] over the news corpora approaches that of a broad-coverage MT system, with one word in ten unknown. These unknown words are principally proper nouns. For the Wikipedia corpus, the coverage is substantially worse. This is due to the fact that this corpus is "dirtier", as well as to the fact that it has contains many more proper nouns.

To measure the performance of the translator we used the Word Error Rate metiric — an edit metric based on the Levenshtein distance (Levenshtein, 1966).

We had two small Kazakh corporas alongside with their postedited translations into Tatar to measure the WER. The first one (2457 words total) was a concatenation of an article from Radio Free Europe / Radio Liberty's Kazakh-language service, an article from Wikipedia, and a simple story used for pedagogical purposes in a workshop on MT for the languages of Rus-

[6] http://www.azattyq.org/
[7] http://kk.wikipedia.org/; kkwiki-20130408-pages-articles.xml.bz2
[8] http://tt.wikipedia.org/; ttwiki-20130205-pages-articles.xml.bz2
[9] http://www.azatlyk.org/
[10] https://apertium.svn.sourceforge.net/svnroot/apertium
[11] Note that the coverage of the vanilla transducers is somewhat higher.

```
<rule>
<match lemma="ақпарат" tags="n.attr"/>
<match lemma="құрал" tags="n.*.px3sp.*"><select lemma="чара"/></match>
</rule>
```

Figure 3: A lexical selection rule that selects чара as the translation of құрал if part of a compound with ақпарат.

(Kazakh) Input	Ол енді ол дыбысты анығырақ ести бастады.
Mor. analysis	ˆОл/ол<det><dem>/ол<prn><dem><nom>/**ол<prn><pers><p3><sg><nom>**$ ˆенді/ен<n><acc>/ен<v><iv><ifi><p3><pl>/ен<v><iv><ifi><p3><sg>/**енді<adv>**$ ˆол/**ол<det><dem>**/ол<prn><dem><nom>/ол<prn><pers><p3><sg><nom>$ ˆдыбысты/**дыбыс<n><acc>**$ ˆанығырақ/анық<adj><comp>/**анық<adj><comp><advl>**/анық<adj><comp><subst><nom>$ ˆести/**есті<v><tv><prc_impf>**$ ˆбастады/баста<v><tv><ifi><p3><pl>/баста<v><tv><ifi><p3><sg>/баста<vaux><ifi><p3><pl> /**баста<vaux><ifi><p3><sg>**$
Mor. disambiguation	ˆОл<prn><pers><p3><sg><nom>$ ˆенді<adv>$ ˆол<det><dem>$ ˆдыбыс<n><acc>$ ˆанық<adj><comp><advl>$ ˆесті<v><tv><prc_impf>$ ˆбаста<vaux><ifi><p3><sg>$ˆ.<sent>$
Lex. transfer (+ selection)	ˆОл<prn><pers><p3><nom>/Ул<prn><pers><p3><nom>$ ˆенді<adv>/инде<adv>/хәзер<adv>$ ˆол<det><dem>/ул<det><dem>$ ˆдыбыс<n><acc>/тавыш<n><acc>$ ˆанық<adj><comp><advl>/анык<adj><comp><advl>$ ˆесті<v><tv><prc_impf>/ишет<v><tv><prc_impf>$ ˆбаста<vaux><ifi><p3><sg>/башла<vaux><ifi><p3><sg>$ˆ.<sent>/.<sent>$
Struct. transfer	ˆУл<prn><pers><p3><nom>$ ˆинде<adv>$ ˆул<det><dem>$ ˆтавыш<n><acc>$ ˆанык<adj><comp><advl>$ ˆишет<v><tv><prc_impf>$ ˆбашла<vaux><ifi><p3><sg>$ˆ.<sent>$
Mor. generation	Ул инде ул тавышны аныграк ишетә башлады.

Table 3: Translation process for the phrase *Ол енді ол дыбысты анығырақ ести бастады* ‘He begins to listen to that sound more carefully’.

Corpus	Tokens	Coverage	stdev
RFERL 2010	3.2M	90.19%	± 0.23%
RFERL 2012	2.9M	89.74%	± 0.59%
Wikipedia	1.2M	80.75%	± 5.23%

(a) Naïve coverage of the Kazakh-Tatar direction

Corpus	Tokens	Coverage	stdev
RFERL 2007-2012	1.2M	82.24%	± 2.88%
New Testament	137K	91.79%	± 1.39%
Wikipedia	128K	%	± %

(b) Naïve coverage of the Tatar-Kazakh direction

Table 4: Naïve coverage of the Kazakh-Tatar system

Corpus	Direction	Tokens	OOV	WER (%)
`devel`	kaz→tat	2457	2	15.19
`test`	kaz→tat	2862	43	36.57

Table 5: Word error rate over two corpora. The `devel` 'development' corpus was used during development to identify possible disambiguation, lexical-selection and transfer rules. The `test` corpus was not seen during development. The OOV column gives the number of out-of-vocabulary (unknown) words.

sia. In addition to postediting the translation, this corpus was run through the morphological transducer and its output was manually disambiguated. All the stems in these texts had been added to the system, and all the rules (Constraint Grammar, lexical selection and transfer) were based on this corpus. Hence we can call it the development corpus. It presents an upper bound on the current performance of the system. Table 5 presents the WER for it.

The second corpus — a collection of only Radio Free Europe / Radio Liberty's Kazakh-language service articles — was used solely for evaluation. It had a similar size — 2862 words. The WER results for it can be seen in the table 5.

Part of the corrections while postediting were due the lack of morphophonoligical rules in Tatar transducer, and were not translation errors as such.

5.1 Error analysis

6 Concluding remarks

To our knowledge we have presented the first ever MT system between Kazakh and Tatar. It has a near-to-production-level coverage, but rather a prototype-level number of rules. Though the impact of this relatively low number of rules on the quality of the translation (compare the WER results for the corpus 1 and corpus 2) looks very promising and suggests the a high-quality translation between morphologically rich agglutinative languages is possible.

We plan to continue development on the pair; the coverage of the system is already quite high, and although we intend to increase it to 95more rules, starting with a Constraint Grammar module. The long-term plan is to integrate the data created with other open-source data for Turkic languages in order to make transfer systems between all the Turkic language pairs. Related work is currently ongoing with Chuvash–Turkish and Turkish–Kyrgyz.

The system is available as free/open-source software under the GNU GPL and the whole system may be downloaded from SVN.[12]

Acknowledgements

The work on this Kazakh-Tatar machine translation system was partially funded by the Google Summer of Code and Google Code-In programs.

References

Altintas, Kemal. 2001a. A morphological analyser for Crimean Tatar. *Proceedings of Turkish Artificial Intelligence and Neural Network Conference*.

Altintas, Kemal. 2001b. Turkish To Crimean Tatar Machine Translation System. Master's thesis, Bilkent University.

Forcada, Mikel L., 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. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Gilmullin, R. A. 2008. The Tatar-Turkish Machine Translation Based On The Two-Level Morphological Analyzer. In *Interactive Systems and Technologies : The Problems of Human-Computer Interaction*, pages 179–186, Ulyanovsk.

Hamzaoğlu, Ilker. 1993. Machine translation from Turkish to other Turkic languages and an implementation for the Azeri language. Master's thesis, Bogazici University.

Karlsson, F., A. Voutilainen, J. Heikkilä, and A. Anttila. 1995. *Constraint Grammar: A language indepen-*

[12] `https://apertium.svn.sourceforge.net/svnroot/apertium/staging/apertium-kaz-tat`

dent 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, Krister, Miikka Silfverberg, Erik Axelson, Sam Hardwick, and Tommi Pirinen, 2011. *HFST—Framework for Compiling and Applying Morphologies*, volume Vol. 100 of *Communications in Computer and Information Science*, pages 67–85.

Tantuğ, A. Cüneyd, Eşref Adalı, and Kemal Oflazer. 2006. Computer analysis of Turkmen language morphology. *Advances in natural language processing, proceedings (LNAI)*, pages 186–193.

Tantuğ, A. Cüneyd, Eşref Adali, and Kemal Oflazer. 2007. A MT system from Turkmen to Turkish employing finite state and Statistical Methods. In *Proceedings of MT Summit XI, Copenhagen, Denmark*.

Tyers, Francis, Jonathan North Washington, Ilnar Salimzyan, and Rustam Batalov. 2012. A prototype machine translation system for Tatar and Bashkir based on free/open-source components. In *Proceedings of the First Workshop on Language Resources and Technologies for Turkic Languages at the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 21.

Washington, Jonathan, Mirlan Ipasov, and Francis Tyers. 2012. A finite-state morphological transducer for Kyrgyz. In Calzolari, Nicoletta (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 23-25. European Language Resources Association (ELRA).

Çağrı Çöltekin. 2010. A freely available morphological analyzer for Turkish. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010)*, pages 820–827.