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

Abstract

This paper presents a bidirectional machine translation system between Kazakh and Tatar, two Turkic languages. Background on the differences between the languages is presented, followed by how the system was designed to handle some of these differences. We provide an evaluation of the system's performance and directions for future work.

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 2 gives a brief description of the two languages; Section 3 gives a short review of some previous work in the area of Turkic–Turkic language translation and compares the grammar of Kazakh and Tatar; 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 several MT systems between Turkic languages has been started (Turkish–Kyrgyz, Azeri–Turkish, Tatar–Bashkir), 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 (Tyers et al., 2012a); 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 except for the Turkish–Azerbaijani pair available through Google Translate.[1] 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 Northwestern) group of Turkic languages. The spoken and written languages share some level of mutual intelligibility to 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 communities 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 har-

[1] http://translate.google.com

mony 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 ‹ъ› 'hard sign' and ‹ь› 'soft sign' 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 ‹ə› in Tatar) follow /j/, a single character is used to represent both: ‹ё›, ‹ю›, and ‹я› respectively.[2]

All of these orthographical conventions present acute 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 do not 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 does not 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.

	[−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

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.

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][3] (where [-eyewitness] is used for cases of both potentially unreliable information and newly discovered information) and [±recent], whereas Tatar has

[2] 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›.

[3] "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.

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 (-Іп-) 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.

	[+recent]	[−recent]
[+reliable]	-DI-	-GAн-
[−reliable]	-Іп-	-GAн екен-

(a) Kazakh past tense morphology

	[−new]	[+new]
[+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

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 '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 minor syntactic differences in Tatar and Kazakh, which include differences in verb valencies in equivalent translations, as well Tatar's reliance on a "true" infinitive that is used in place of various verbal noun and verb adverb forms in Kazakh.

An example of a difference in verb valencies is with the expression corresponding to "to like to do something" in Kazakh and Tatar. In Kazakh, the verb ұна is used, as shown in example (1a), where the subject "I" in English is expressed through a dative experiencer in Kazakh and the gerundive "writing dictionaries" is the grammatical subject. Tatar, on the other hand, uses a verb whose arguments correspond to the arguments of "to like" in English, as shown in example (1b), where the first person pronoun is in nominative case as the grammatical subject and the infinitival verb phrase is the grammatical direct object.

(1) a. Маған сөздік түзген
маған сөздік түз-GAн
1P.SG.DAT dictionary compile-GER
ұнайды.
ұна-Е-дI
like-AOR-3P.SG
'I like writing dictionaries.'

b. Мин сүзлек төзергә
мин сүзлек төз-IргA
1P.SG dictionary compile-INF
яратам.
ярат-Е-м
like-PRES-1P.SG
'I like writing dictionaries.'

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 in (2-3).

(2) a. Мен үйге қайтуым керек.
мен үй-GA қайт-у-Ім 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.'

(3) a. Айгүл оны табуға әрекет
Айгүл о-NI тап-у-GA әрекет
Aygül 3P.SG-ACC find-GER-DAT effort
жасап жүр.
жаса-Ip жүр
make-PRC PROG
'Aygül is trying to find him.'

b. Айгөл аны табарга тырыша.
Айгөл а-NI тап-АргА тырыш-Е
Aygöl 3P.SG-ACC find-INF тырыш-PRES
'Aygöl is trying to find him.'

As shown in (4), the Tatar infinitive also corresponds to a verbal adverb form in Kazakh.

(4) a. Мен сенімен сөйлескелі
мен сен-мен сөйле-с-GAлI
I you-INST talk-COOP-VADV
келдім.
кел-DI-м.
come-IFI-1P.SG
'I came to speak with you.'

b. Мин синең белән сөйлешергә
мин син-Iң белән сөйле-ш-IргА
I you-GEN with talk-COOP-INF
килдем.
кил-DI-м.
come-IFI-1P.SG
'I came to speak with you.'

This example also demonstrates the correspondence of the Kazakh intstrumental case -Мен to the Tatar postposition белән 'with', which are cognate structures; while their phonology and orthographic standards differ, they are largely parallel in use.

4 System

The system is based on the Apertium machine translation platform (Forcada et al., 2011).[4] 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[5] (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] http://www.apertium.org
[5] http://www.fsf.org/licensing/licenses/gpl.html

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 (Tyers et al., 2012b) 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 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 (Çöltekin,

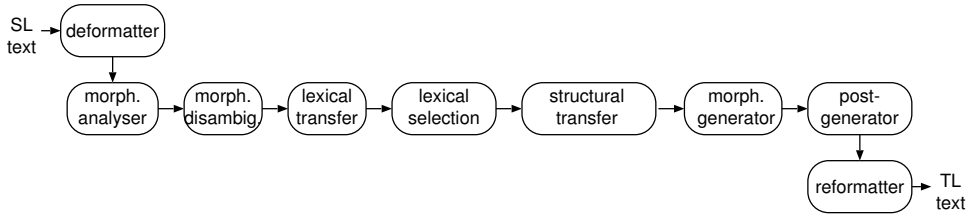

Figure 1: The pipeline architecture of the Apertium system.

2010; Altintas, 2001a; Tantuğ et al., 2006; Washington et al., 2012; Tyers et al., 2012a), 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 takes a harmonised low vowel in the infinitive. Class membership cannot be predicted based on any phonological criteria and is simply a lexical property of any given verb. This was implemented in lexc with two similar sets of 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. These two sets of continuation lexica are otherwise the same.

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 unrecognised by the morphological analyser into Tatar). Some toponyms and other proper names were translated semi-automatically by looking up links in Wikipedia (Tyers and Pienaar, 2008); 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.[6]

The output of each morphological analyser is highly ambiguous, measured at around 2.4 morphological analyses per form for Kazakh. The goal of the CG rules is to select the correct analysis when there are multiple analyses. As of r43664, the ambiguity was down to 1.4 analyses per form.

One reason for the still high level of morphological ambiguity is a series of affixes in both languages which can each be analysed as some combination of verbal nouns, verbal adjectives, substantitivised verbal adjectives (i.e., verbal adjectives with a null modified noun), and even finite forms.[7]

Given the similarity of Kazakh and Tatar, this sort of ambiguity may often be passed from one language to the other and not lead to many translation errors. While disambiguating between these analyses would be crucial for e.g., a Turkic-to-English system, we have not yet put much effort into developing CG rules to deal with such ambiguity.

4.5 Lexical selection rules

While many lexical items have a similar range of meaning, lexical selection can sometimes be problematic be-

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

[7] Despite the fact that the various suffixes in this category pattern differently and do not form a single natural class, most grammars of the languages label all them as simply “gerunds” or “participles”.

```
<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

```
<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.

tween Kazakh and Tatar.

For example (see Fig. 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 Fig. 3.

Likewise, the Kazakh word топ can be translated to Tatar as туп 'ball' (sometimes доп in Kazakh), and as 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.

Tatar has separate words for '[physical] life' (гомер) and 'life [as a human condition]' (тормыш), whereas Kazakh only has one word (өмір). The lexical selection rule provided in Fig. 4 chooses the latter Tatar translation after the adjective рухани 'spiritual'.

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

```
<rule>
  <match lemma="рухани" tags="adj"/>
  <match lemma="өмір" tags="n.*">
      <select lemma="тормыш"/>
  </match>
</rule>
```

Figure 4: A lexical selection rule that selects тормыш for өмір if preceded by the word рухани.

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	81.36%	± 1.48%

(b) Naïve coverage of the Tatar-Kazakh direction

Table 3: Naïve coverage of the Kazakh-Tatar system

(Kazakh) Input	Ол енді ол дыбысты анығырақ ести бастады.
Mor. analysis	^Ол/ол<det><dem>/ол<prn><dem><nom>/**ол<prn><pers><p3><sg><nom>**$ ^енді/ен<n><acc>/ен<v><iv><ifi><p3><pl>/ен<v><iv><ifi><p3><sg>/**енді<adv>**$ ^ол/**ол<det><dem>**/ол<prn><dem><nom>/ол<prn><pers><p3><sg><nom>$ ^дыбысты/**дыбыс<n><acc>**$ ^анығырақ/анық<adj><comp>/**анық<adj><comp><advl>**/анық<adj><comp><subst><nom>$ ^ести/**есті<v><tv><prc_impf>**$ ^бастады/баста<v><tv><ifi><p3><pl>/баста<v><tv><ifi><p3><sg>/баста<vaux><ifi><p3><pl> /**баста<vaux><ifi><p3><sg>**$
Mor. disambiguation	^Ол<prn><pers><p3><sg><nom>$ ^енді<adv>$ ^ол<det><dem>$ ^дыбыс<n><acc>$ ^анық<adj><comp><advl>$ ^есті<v><tv><prc_impf>$ ^баста<vaux><ifi><p3><sg>$^.<sent>$
Lex. transfer (+ selection)	^Ол<prn><pers><p3><nom>/Ул<prn><pers><p3><nom>$ ^енді<adv>/инде<adv>/хәзер<adv>$ ^ол<det><dem>/ул<det><dem>$ ^дыбыс<n><acc>/тавыш<n><acc>$ ^анық<adj><comp><advl>/анык<adj><comp><advl>$ ^есті<v><tv><prc_impf>/ишет<v><tv><prc_impf>$ ^баста<vaux><ifi><p3><sg>/башла<vaux><ifi><p3><sg>$^.<sent>/.<sent>$
Struct. transfer	^Ул<prn><pers><p3><nom>$ ^инде<adv>$ ^ул<det><dem>$ ^тавыш<n><acc>$ ^анык<adj><comp><advl>$ ^ишет<v><tv><prc_impf>$ ^башла<vaux><ifi><p3><sg>$^.<sent>$
Mor. generation	Ул инде ул тавышны аныграк ишетә башлады.

Table 4: Translation process for the Kazakh phrase *Ол енді ол дыбысты анығырақ ести бастады* 'He begins to listen to that sound more carefully'.

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 (RFERL)'s Kazakh-language service,[8] as well as a recent dump of Wikipedia's articles in Kazakh[9] were used.

For Tatar, a dump of articles from the Tatar Wikipedia,[10] a translation of the New Testament, and content from RFERL's Tatar-language service[11] from early 2007 to early 2012 were used for testing.

The versions of the transducers tested were `r43595` from the Apertium SVN.[12] 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 3, the naïve coverage of the Kazakh-Tatar MT system[13] over the news corpora approaches that of a broad-coverage MT system, with one word in ten unknown. The coverage over the Wikipedia corpus is substantially worse, due to the fact that this corpus is "dirtier": it contains orthographical errors, *wiki* code, repetitions, as well as quite a few proper nouns.

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; OOV is the number of out-of-vocabulary (unknown) words.

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

We had two small Kazakh corpora along with their postedited translations into Tatar to measure the WER. The first one (2,457 words total) was a concatenation of an article from RFERL's Kazakh-language service, an article from Wikipedia, and a simple story used for pedagogical purposes in a workshop on MT for the lan-

[8] http://www.azattyq.org/

[9] http://kk.wikipedia.org/; `kkwiki-20130408-pages-articles.xml.bz2`

[10] http://tt.wikipedia.org/; `ttwiki-20130205-pages-articles.xml.bz2`

[11] http://www.azatlyk.org/

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

[13] The coverage of the vanilla transducers is slightly higher.

form	current	expected
ди<v><tv><pres><p3><sg>	дия	ди
йөр<v><iv><pres><p3><sg>	йөрə	йөри
кияү<n><px3sp><nom>	кияүе	кияве
укы<v><iv><ger><nom>	укыу	уку

Table 6: Examples of some phonological problems in the Tatar transducer.

guages of Russia. In addition to postediting the translation, we ran this corpus through the morphological transducer and manually disambiguated its output. All the stems in these texts were added to the system, and all the rules (CG, lexical selection, and transfer) were based on this corpus. This "development" corpus presents an upper bound on the current performance of the system. The testing corpus, used solely for evaluation, was comprised of articles from RFERL's Kazakh-language service, and had a similar size (2,862 words) to the development corpus. Table 5 presents the WER for both corpora.

Some of the corrections that were made as part of postediting were not translation errors as such, but were instead due to the lack of morphophonological rules in the Tatar transducer. A few examples, which include irregular verbs (ди and йөр) and are otherwise orthographical corner cases, are given in table 6. Aside from the irregular behaviour of йөр, which would need its irregularities implemented through special phonotactics in lexc, these issues are all shortcomings of our phonological layer, implemented in twol.

The majority of remaining errors are mostly due to mistakes and gaps in the Tatar morphophonology component, lack of transfer rules to handle some Kazakh compounds, and disambiguation errors.

6 Concluding remarks

To our knowledge we have presented the first ever MT system between Kazakh and Tatar. It has near-production-level coverage, but is rather prototype-level in terms of the number of rules. Although the impact of this relatively low number of rules on the quality of translation is extensive (cf., the difference in WER between the development and testing corpora), the outlook is promising and the current results suggest that 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 95% on the corpora we have, the main work will be improving the quality of translation by adding more rules, starting with the 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.[14]

Acknowledgements

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

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 independent system for parsing unrestricted text*. Mouton de Gruyter.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet*

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

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, F. M. and J. A. Pienaar. 2008. Extracting bilingual word pairs from wikipedia. In *Proceedings of the SALTMIL Workshop at the Language Resources and Evaluation Conference, LREC2008*, pages 19–22.

Tyers, Francis, Jonathan North Washington, Ilnar Salimzyan, and Rustam Batalov. 2012a. 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.

Tyers, Francis M., Felipe Sánchez-Martínez, and Mikel L. Forcada. 2012b. Flexible finite-state lexical selection for rule-based machine translation. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 213–220, Trento, Italy, May.

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).

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