Finite-state morphological transducers for three Kypchak languages

Jonathan North Washington[†], **Ilnar Salimzyanov**[‡], **Francis M. Tyers**[⋆]

[†]Departments of Linguistics and Central Eurasian Studies
Indiana University
Bloomington, IN 47405 (USA)
jonwashi@indiana.edu

[‡]Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Stuttgart (Germany)
ilnar@ilnar

[⋆]Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03071 Alacant (Spain)
ftyers@dlsi.ua.es

Abstract

Hargle, bargle.

Keywords: Kazakh, Tatar, Kumyk, morphology, transducer

1. Introduction

This paper describes the development of morphological transducers for three closly related languages: Kazakh, Tatar, and Kumyk.

The transducers for these languages

2. Languages

The three languages for which transducers were developed belong to the Northwestern branch of Turkic, which is often referred to as the Kypchak branch. This branch can be divided into three subbranches. Kumyk is a member of the Western Kypchak group, Tatar is a member of the Northern Kypchak group, and Kazakh is a member of the Southern Kypchak group (Johanson, 2006, 82-83). As such, each of these three languages represents a different one of the three branches of Kypchak. The geographic distribution of the languages is shown in map 1.

These languages have different amounts of linguistic influence from other Turkic branches (e.g., moderate Oghuz (SE) influence in the Western group, slight Oghuz influence in the Northern group) and from Mongolic languages (moderate influence on the Southern group, lighter in the other groups), and all have heavy influence from Persian.

2.1. Kazakh

Kazakh /qɑzɑq/ is spoken primarily 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. Estimates of the total number of speakers range from 8 million (Lewis et al., 2013) to 11 million (Nationalencyklopedin, 2013) people.

2.2. Tatar

Tatar /tɒtɑr/ is spoken in and around Tatarstan by approximately 5.4 million people (Lewis et al., 2013). It is co-official with Russian in Tatarstan — a republic within Russia. A majority of native speakers of both languages are bilingual in Russian.

2.3. Kumyk

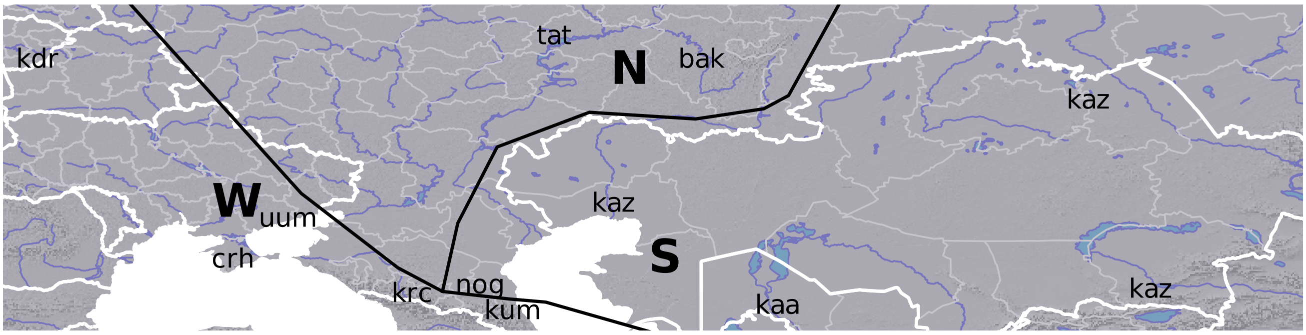
Kumyk /qumuq/ is spoken in Dagestan, a Republic of the Russia Federation, where it is co-official with a number of other languages of Dagestan (Lewis et al., 2013). There are approximately 430 thousand speakers (Lewis et al., 2013).

3. Background

3.1. Morphological transducers, previous work

The objective of a morphological transducer is twofold: firstly to take surface forms (e.g., алдым) and generate all possible lexical forms, and secondly to take lexical forms (e.g., ал`<v><tv><ifi><p1><sg>`, алд`<n><px1sg><nom>`, etc.) and generate one or more surface forms. Since they are implemented as finite-state automata, they are reversible by default.

The transducers were designed based on the Helsinki Finite State Toolkit (**?**) which is 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,



Map 1: The three subbranches of Kypchak (N, S, W), roughly divided with black lines, showing the geographic distribution of the three languages for which transducers were developed (tat, kaz, kum). Language codes are from ISO 639-3.

such as Turkish (Çöltekin, 2010), Crimean Tatar (Altintas, 2001), Turkmen (Tantuğ et al., 2006), and Kyrgyz (Washington et al., 2012), and is available under a free/open-source licence.

The authors learned of another Kazakh morphological transducer in existence (Бекманова & Махимов, 2013) only after production-ready version of our transducer was released. We have not yet been able to evaluate this system or compare it to ours.

Creating a morphological transducers in the above-mentioned formalisms simply involves encoding linguistic knowledge about the language in the formalisms. The lexc and twol formalisms resemble linguistic formalisms, allowing the coders to work with abstractions resembling linguistic categories such lexemes, morphemes, phonemes, and even archiphonemes—as opposed to a raw FST, where input characters are translated to output characters along a graph.

3.2. Description

The transducers are available / under development in apertium's subversion repository,[1] in the directories apertium-kaz, apertium-tat, and apertium-kum. The revision of the entire subversion repository that the numbers (stem counts, evaluation, etc.) in this paper represent is `r48137`.

The tagset consists of 127 separate tags, 19 covering the main parts of speech (noun, verb, adjective, adverb, postposition, etc.) and 108 covering morphological subcategorisation for e.g. case, number, person, possession, transitivity, tense-aspect-mood, etc. The tags are represented as multicharacter symbols, between less than '<' and greater than '>' symbols. The tagset is quite extensive and still not entirely stabilised, as such a full listing is not included here. However, the tags are listed in the source code of the transducer,[2] along with comments describing their usage.

4. Methodology

4.1. Development effort

The three transducers discussed in this paper are for Kazakh, Tatar, and Kumyk. The Kazakh and Tatar transducers were originally created as part of an experimental Kazakh-Tatar machine translation system in December of 2010. The Kazakh transducer was expanded during Google Code-In 2010 and 2011, and the Tatar transducer was expanded as part of a prototype Tatar and Bashkir machine translation system (Tyers et al., 2012). The Kazakh-Tatar machine translation system, along with the two transducers, was expanded to production-level quality as part of a Google Summer of Code project in 2012 (Salimzyanov et al., 2013).

The Kumyk transducer was developed starting at the beginning of October, 2013 as experiment to see how difficult it would be to extend lessons learned from the development of the Tatar and Kazakh transducers to a related language. This paper explores some of these lessons and how the development of the Kumyk transducer benefited from knowledge gained from the development of the Tatar and Kazakh transducers.

The morphotactics of Turkic languages are complex enough that even a linguist who is fluent in the language and has a good linguistic understanding of it may not understand how exactly all morphemes combine. Native speakers educated about the morphology of their languages also do not have an explicit knowledge of the complete morphotactics. Hence it often becomes necessary to use fieldwork methodology to elicit the full extent of the morphotactics, be this a linguist with no to little knowledge of a Turkic language working with a native speaker, or a native speaker who

[1] https://svn.code.sf.net/p/apertium/svn/languages/

[2] no url needed?

Part of speech	Number of stems		
	Kazakh	Tatar	Kumyk
Noun	-	-	-
Verb	-	-	-
Adjective	-	-	-
Proper noun	-	-	-
Adverb	-	-	-
Numeral	-	-	-
Conjunction	-	-	-
Postposition	-	-	-
Pronoun	-	-	-
Determiner	-	-	-
Total:	-	-	-

Table 1: Number of stems in each of the categories

Language	Corpus	Words	Coverage
Kazakh	Wikipedia 2013	-	-
	RFE/RL 2010	3.2M	-
	Bible	577K	-
	Average	-	90.5%
Tatar	Wikipedia 2013	128K	-
	News 2005–2011	4.6M	-
	New Testament	137K	-
	Average	-	89.0%
Kumyk	Yoldaš	287K	-
	New Testament	154K	-
	Average	-	90.1%

Table 2: Corpora used for naïve coverage tests

understands the extent of what knowledge is necessary to encode in the transducer. When there is no native speaker of a particular language available, the authors have found that information previously encoded about a closely related language or the intuitions of a speaker of a closely related language may be combined with the use of textual corpora to "elicit" information about the morphotactics of a language. Depending on the contents of corpus and chance, this may not result in a completely accurate model, but it is possible to be thorough.

The Kazakh morphotactics were originally developed based on the Kyrgyz transducer, which was coauthored by a linguist who is fluent in and has a good linguistic knowledge of Kyrgyz together with a native speaker of Kyrgyz, and in consultation with another native speaker of Kyrgyz. The initial developer of the Kazakh morphotactic was the same linguist, who is fluent in and has a good linguistic understanding Kazakh. The morphotactics of Tatar were developed for the most part by a native Tatar speaker, who also worked to polish off the morphotactics of the Kazakh transducer.

As the authors found it difficult to locate native speakers of Kumyk, the morphotactics of the Kumyk transducer were developed based on the existing encoded morphotactics of Kazakh and Tatar (and occasionally Kyrgyz), with consultation of corpora, as described above. A dictionary (Бamмaтов, 1960) and a grammar (Ольмесов, 2000) of Kumyk were also consulted as needed.

4.2. Statistics

5. Evaluation

We have evaluated the morphological analysers in two ways. The first was by calculating the naïve coverage[3] and mean ambiguity on freely available corpora. The mean ambiguity measure was calculated by performing an evaluation of precision and recall on some smaller, hand-validated test sets.

5.1. Corpora

We tested the coverage of the Kazakh and Tatar analysers over three separate domains: encyclopaedic text,[4] news,[5] and religion.[6] As there is currently no wikipedia in Kumyk, we tested only news and religion.[7]

The coverage of each transducer over the various corpora is shown in table 2.

[3]Naïve coverage refers to the percentage of surface forms in a given corpora that receive at least one analysis. Forms counted by this measure may have other analyses which are not delivered by the transducer.

[4]The following wikipedia dumps were used: kkwiki-20131006-pages-articles.xml.bz2, FIXME.

[5]All content from RFE/RL (http://www.azattyq.org/) for 2010 was used for Kazakh, as well as all content from 2005 to 2011 on http://tat.tatar-inform.ru for Tatar.

[6]We used a Kazakh bible translation available from https://kkitap.net/ and a Tatar translation of the New Testament available from http://ibt.org.ru

[7]The bible corpus is from http://ibt.org.ru/ and the news corpus consists of all Kumyk from [8].

Language	Precision	Recall
Kazakh	-	-
Tatar	-	-
Kumyk	-	-

Table 3: Precision and recall

6. Future work

7. Conclusions

Acknowledgements

We would like to thank the Google Code-in (2011) for supporting the development of the Kazakh transducer, and in particular the effort by Nathan Maxson. We would also like to thank the Google Summer of Code (2012) for supporting the development of both the Kazakh and the Tatar transducers.

References

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

Johanson, Lars (2006). History of Turkic. In Lars Johanson & Éva Á. Csató (Eds.), *The Turkic Languages*, New York: Routledge, chap. 5, pp. 81–125.

Lewis, M. Paul, Simons, Gary F., & Fennig, Charles D. (Eds.) (2013). *Ethnologue: Languages of the World*. Dallas, Texas: SIL International, seventeenth edn. http://www.ethnologue.com.

Linden, Krister, Silfverberg, Miikka, Axelson, Erik, Hardwick, Sam, & Pirinen, Tommi (2011). *HFST—Framework for Compiling and Applying Morphologies*, vol. Vol. 100 of *Communications in Computer and Information Science*, pp. 67–85. ISBN 978-3-642-23137-7.

Nationalencyklopedin (2013). kazakiska. In *Nationalencyklopedin*. http://www.ne.se/.

Salimzyanov, Ilnar, Washington, Jonathan North, & Tyers, Francis M. (2013). A free/open-source Kazakh-Tatar machine translation system.

Tantuğ, A.C., Adalı, E., & Oflazer, K. (2006). Computer analysis of Turkmen language morphology. *Advances in natural language processing, proceedings (Lecture notes in artificial intelligence)*, pp. 186–193.

Tyers, Francis, Washington, Jonathan North, Salimzyan, Ilnar, & Batalov, Rustam (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.

Washington, Jonathan North, Ipasov, Mirlan, & Tyers, Francis M. (2012). A finite-state morphological analyser for Kyrgyz.

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

Бамматов, З. З. (1960). *Русско-кумыкский словарь*. Москва: Государственное издательсвто иностранных и национальных словарей.

Бекманова, Г. Т. & Махимов, А. (2013). Графематический и моргологический анализатор Казахского языка. pp. 192–200.

Ольмесов, Нураммат Хайруллаевич (2000). *Сопоставительная грамматика кумыкского и русского языков*. Махачкала: ИПЦ ДГУ.