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.

These languages 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 **??**.

In a linguistic sense, these languages have different amounts of 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.

Бекманова & Махимов (2013)

The transducers for these languages

2. Languages

2.1. Kazakh

Kazakh /q□z□q/ 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 (**?**) to 11 million (**?**) people.

2.2. Tatar

Tatar /t□t□r/ is spoken in and around Tatarstan by approximately 5.4 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.

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 (**?**). There are approximately 430 thousand speakers (**?**).

Бамматов (1960) Ольмесов (2000)

3. Background

3.1. Morphological transducers

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, such as

Turkish (**?**), Crimean Tatar (**?**), Turkmen (**?**), and Kyrgyz (Washington et al., 2012), and is available under a free/open-source licence.

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.

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.

3.2. Description

The transducers are available / under development at `https://svn.code.sf.net/p/apertium/svn/languages/`, 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,[1] along with comments describing their usage.

[1] `https://apertium.svn.sourceforge.net/`

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	-
	RFE/RL 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

4. Methodology

4.1. Development effort

4.2. Statistics

5. Evaluation

We have evaluated the morphological analysers in two ways. The first was by calculating the naïve coverage[2] and mean ambiguity on freely available corpora. The second was by performing an evaluation of precision and recall on some smaller, hand-validated test sets.

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

Table 3: Precision and recall

5.1. Corpora

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

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

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

Tyers, Francis, Washington, Jonathan North, & an d Rustam Batalov, Ilnar Salimzyan (2012). A prototype machine translation system for Tatar and Bashkir bas ed on free/open-source components. In *Proceedings of the First Workshop on Language Resources and Tech nologies 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.

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.

svnroot/apertium/branches/apertium-kir/

[2]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.