

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 closely 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 sub-branches. 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.

Washington et al. (2012) Salimzyanov et al. (2013)
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. Methodology

3.1. Development effort

3.2. Statistics

4. Evaluation

We have evaluated the morphological analysers in two ways. The first was by calculating the naïve coverage¹ 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.

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	Yoldaş	287K	-
	New Testament	154K	-
	Average	-	90.1%

Table 2: Corpora used for naïve coverage tests

4.1. Corpora

5. Future work

6. Conclusions

Acknowledgements

We would like to thank the Google Code-in (2011) for supporting the development of the Kazakh trans-

ducer, 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.
- 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 издательство иностранных и национальных словарей.
- Бекманова, Г. Т. & Махимов, А. (2013). Графематический и морфологический анализатор Казахского языка. pp. 192--200.
- Ольмесов, Нурамат Хайруллаевич (2000). *Сопоставительная грамматика кумыкского и русского языков*. Махачкала: ИПЦ ДГУ.

¹Naïve 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.

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

Table 3: Precision and recall