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

The Northwestern branch of Turkic is often referred to as the Kypchak branch, and 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 south Kypchak group (Johanson, 2006, 82-83). The geographic distribution of the languages is shown in map ??.

Washington et al. (2012) Salimzyanov et al. (2013) Бекманова and Махимов (2013)

2. Languages

2.1. Kazakh

2.2. Tatar

2.3. Kumyk

Бамматов (1960)

3. Methodology

4. Evaluation

5. Future work

6. Conclusions

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 издательсвто иностранных и национальных словарей.

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

Language	-
Kazakh	-
Tatar	-
Kumyk	-

Table 1: Naïve coverage

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

Table 2: Precision and recall

Corpus	Words	Coverage
wikipedia	850K	-
Äwezov	155K	-
RFERL 2010	3.2M	-
bible	577K	-
quran	107K	-

Table 3: Corpora used for coverage of Kazakh

Language	Corpus	Words	Coverage
Kazakh	wikipedia 2011	850K	-
	Äwezov	155K	-
	RFERL 2010	3.2M	-
	bible	577K	-
	quran	107K	-
	average	-	90.5%
Tatar	wikipedia 2013	128K	-
	news 2005-2011	4.6M	-
	new testament	137K	-
	quran	165K	-
	Aytmatov	5K	-
	average	-	89.0%
Kumyk	yoldash	287K	-
	new testament	154K	-
	book of Genesis	28K	-
	average	-	88.0%

Table 4: Corpora used for coverage tests

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