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 & Махимов (2013)

2. Languages

2.1. Kazakh

2.2. Tatar

2.3. Kumyk

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 coverage[1] 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	Yoldaš	287K	-
	New Testament	154K	-
	Average	-	90.1%

Table 2: Corpora used for naïve coverage tests

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

Table 3: Precision and recall

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 transducer and 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 издательсвто иностранных и национальных словарей.

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

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

[1]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.