3rd International Conference on Computer Processing
in Turkic Languages (TURKLANG 2015)

Towards a free/open-source universal-dependency treebank for Kazakh

Francis M. Tyers[a] **and Jonathan Washington**[b]

[a] HSL-fakultehta, UiT Norgga árktalaš universitehta, N-9015 Tromsø, Norway
[b] Departments of Linguistics and Central Eurasian Studies, Indiana University, Bloomington, IN 47405, USA

Abstract

This article describes the first steps towards a free/open-source dependency treebank for Kazakh based on universal dependency (UD) annotation standards. The treebank contains 302 sentences and is based on texts from a range of open-source and public domain sources. This ensures its free availability and extensibility. Texts in the treebank are first morphologically analysed and disambiguated and then annotated manually for dependency structure.

Keywords: Kazakh; treebank; dependency grammar; universal dependency

1 Introduction

?, ?, ?, ?

Document	Description	Sentences	Tokens	Avg. length
UN Declaration on Human Rights	Legal text on human rights	25	409	16.3
Phrasebook	Phrases from Wikitravel	37	205	5.5
Жиырма Бесінші Сөз	Philosophical text	34	525	15.4
Өлген қазан	Folk tale	8	140	17.5
Ер Төстік	Folk tale	23	203	8.8
Азамат қайда?	Story for language learners	48	434	9.0
Футболдан әлем чемпионаты 2014	Wikipedia article (2014 World Cup)	14	246	17.5
Иран	Wikipedia article (Iran)	111	1562	14.0
Радиан	Wikipedia article (Radian)	2	17	8.5
		302	3741	12.3

Table 1: Composition of the corpus. The corpus covers a range of genres and text types from free and public-domain sources.

Table 2: Universal dependency label set

2 Background

2.1 Kazakh

2.2 Treebanks

3 Methodology

3.1 Corpus

3.2 Preprocessing

Preprocessing the corpus consists of running the text through the Kazakh morphological analyser (**?**), which also performs tokenisation of multiword units based the longest match left-to-right. Tokenisation for Kazakh is a non-trivial task, and so we do not simply take space as a delimiter. The morphological analyser returns all the possible morphological analyses for each word based on a lexicon of around 20,000 lexemes. After tokenisation and morphological analysis, the text is processed with a constraint-grammar based disambiguator for Kazakh consisting of 113 rules which remove inappropriate analyses in context. This reduces the average number of analyses per word from around 3.4 to around 1.7.

3.2.1 Tokens and words

4 Annotation guidelines

4.1 Copula

4.2 Coordination

4.3 Complex nominals

4.4 Non-finite clauses

5 Evaluation

6 Future work

7 Conclusions

Acknowledgements