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

250-300-word "extended abstract in English"

Keywords: Kazakh; treebank; dependency grammar; universal dependency

1   Introduction

?, ?, ?, ?

2   Background

2.1   Kazakh

2.2   Treebanks

3   Methodology

3.1   Corpus

3.2   Preprocessing

?

Document	Description	Sentences	Tokens	Avg. length
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
Азамат қайда?	Story for children and 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

Table 1: Composition of the corpus. The corpus covers a range of genres and text types.

Table 2: Universal dependency label set

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