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

This article describes the work towards the development of a dependency treebank for Kazakh, a Turkic language spoken in Central Asia. Despite its status as a *core* Turkic language, little computational-linguistic research has been published on syntactic parsing for Kazakh. A valuable resource in the study of syntactic parsing is a treebank—a corpus of parsed text containing gold-standard syntactic annotation.

Freely available treebanks exist for many languages, such as large languages like Finnish (x,y) and Polish (xxx) and smaller languages like Irish (Lynn et al., 2012). To our knowledge only one treebank exists for another language, Turkish (Oflazer et al., 2003), which is unfortunately not freely available.

In building our treebank we take advantage of existing work done on tokenisation, morphological analysis and part-of-speech tagging for Kazakh. We also take a pragmatic and iterative view of development of the treebank, in line with recent work on cross-linguistic parsing with universal dependencies (Marneffe et al., 2014).

The remainder of the paper is organised as follows. Section **??** gives some background linguistic information on Kazakh, and outlines some special challenges in parsing Kazakh. In Section **??** we describe the corpus itself and methodology used in annotating it. Section **??** gives a sketch of some decisions we have made with respect to annotation guidelines, referring back to the discussion in Section **??**. For reasons of space, these guidelines are not complete, but present a subset of guidelines which are of particular interest. A small experiment in statistical dependency parsing using the corpus is presented in Section **??**, and in Sections **??** and **??** we give perspectives for future work and some concluding remarks.

2 Background

2.1 Kazakh

Kazakh (қазақ тілі), a Turkic language of Central Asia, is spoken by around 13 million people in Kazakhstan, China, Mongolia, and adjacent areas (Lewis et al., 2015). While works like Балақаев et al. (1954) provide decent syntactic descriptions of the language, there is little to no work on the syntax of Kazakh within modern theoretical syntactic frameworks (though work on related languages, such as Turkish, exists).

As an agglutinative language with rich morphology and agreement phenomena, Kazakh presents some interesting challenges for computational syntax. A morphological transducer has already been designed (Washington et al., 2014), which implements a number of decisions about how various inflectional morphemes affect the part of speech and syntactic function of various word forms. These decisions include how the relations "case" morphemes create between syntactic elements work, as well as a number of processes of "zero derivation".

All of the traditionally defined "cases" in Kazakh have a variety of uses. The morphemes that mark the accusative and genitive cases are only used when the noun phrase is semantically definite (e.g., үйді *the house* `acc`, үйдің *of the house* `gen`). This means that indefinite accusative and genitive as well as nominative noun phrases are all unmarked for case, and hence are all ambiguous (e.g., үй *house*). The Kazakh transducer marks these all as `<nom>`, but their use may be disambiguated. Besides identifying the semantic role of an unmarked noun phrase, there are various agreement properties that make it clear what "case" the noun phrase "truly" receives: genitive noun phrases must have a corresponding noun phrase with possessive morphology that agrees in person with the genitive noun phrase; and truly nominative noun phrases must have a corresponding predicate that agrees in person with the nominative noun phrase—e.g., a verb or a copula.

Another category that may also be non-overt is the copula. In the present tense, all but third person copulas have realised forms (e.g., Мен үйде□□□. *I'm at home.*), but the third person copula does not surface (e.g., Ол үйде. *S/he's at home.*).

Another type of "zero-derivation" involves the use of adjectives and nouns. In Kazakh, adjectives are all attributive, but many may also be used either nominally (i.e., as nouns) or adverbially (i.e., as adverbs), or both. The morphological transducer internally provides a series of classes; depending on class membership, an adjective may receive the readings `<adj><advl>` or `<adj><subst>`, besides the default attributive reading of `<adj>`. If an adjective may be used substantivally, then it will receive a range of substantive morphology as well. Similarly, many nouns may be used attributively, besides the default substantive reading. Besides the nominative (and indefinite accusative/genitive) reading of `<nom>`, the transducer provides an attributive reading for most nominals of `<nom><attr>`.

How each of these issues bears on a dependency analysis will be discussed in section **??**.

2.2 Treebanks

3 Methodology

3.1 Corpus

3.2 Preprocessing

Preprocessing the corpus consists of running the text through the Kazakh morphological analyser (Washington et al., 2014), 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.

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
Қожанасырдың тойға баруы	Folk tale from Wikisource	8	140	17.5
Ер Төстік	Folk tale from Wikisource	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.

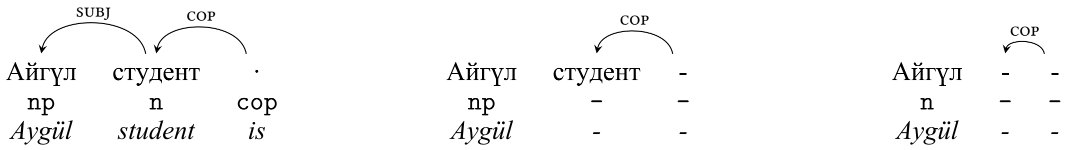
3.2.1 Tokens and words

4 Annotation guidelines

Label	Description

Table 2: Universal dependency label set

4.1 Copula



(a) xxx

(b) xxx.

(c) xxx.

Figure 1: Dependency trees of copula constructions.

4.2 Coordination

4.3 Complex nominals

There are different ways in which two nominals may occur together to act as a single nominal. Compounds are formed by an attributive nominal (indistinguishable from the bare / nominative form, but tagged with `<attr>`) preceding another nominal, as shown in 3a. An indefinite genitive construction is formed by an indefinite genitive nominal (indistinguishable from the bare / nominative form, and tagged with `<nom>`) preceding a nominal that has third-person possessive morphology, as shown in 3b. A definite genetive construction is formed by a genitive-marked nominal (tagged `<gen>`) precding a nominal that has third-person possessive morphology, as shown in 3c.

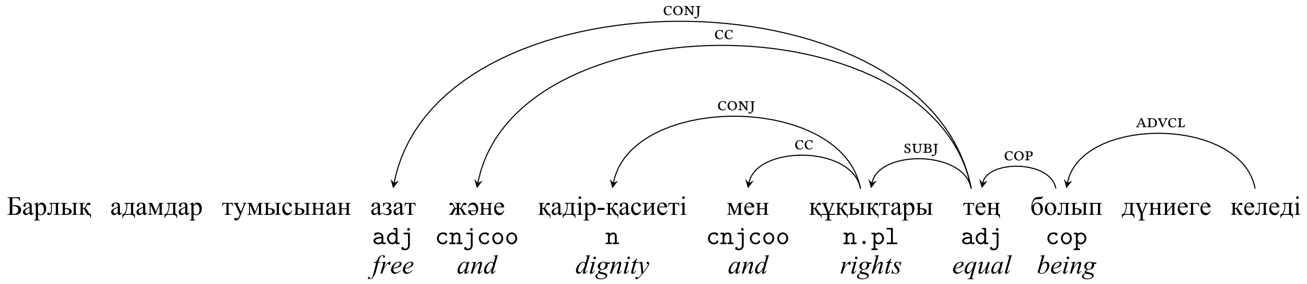
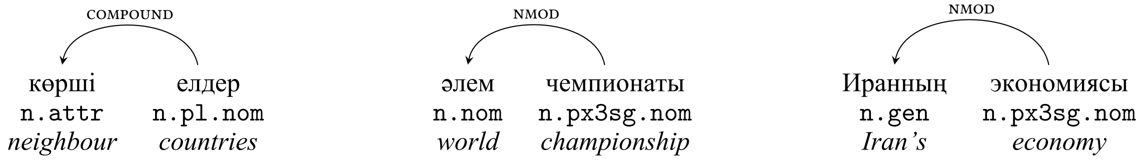

CONJ
CC
CONJ
CC
SUBJ
COP
ADVCL
Барлық адамдар тумысынан азат және қадір-қасиеті мен құқықтары тең болып дүниеге келеді
adj cnjcoo n cnjcoo n.pl adj cop
free and dignity and rights equal being

Figure 2: Coordination



(a) Compounding: an attributive nominal depending on a nominal.

(b) Indefinite genitive: An indefinite genitive depending on a nominal.

(c) Definite genitive: A definite genitive depending on a nominal.

Figure 3: Dependency trees of complex nominal relations.

As seen in the graphs, the compound relationship of an attributive nominal depending on another nominal is labelled COMPOUND, and genitive relations are considered NMOD, regardless of whether there is a definite or indefinite genitive construction.

4.4 Non-finite clauses

5 Evaluation

In order to test the utility of the treebank in a real-world setting, we trained and evaluated a number of models using the popular MaltParser tool (Nivre et al., 2007). MaltParser is a toolkit for data-driven dependency parsing, it can learn a parsing model from treebank data and apply this model to parse unseen sentences. The parser has a large number of options and parameters that need to be optimised. To select the best parser configuration we relied on MaltOptimiser (Ballesteros et al., 2015). The optimiser was run separately for each of the model configurations.

As the treebank takes advantage of the new tokenisation standards in the CoNLL-U format, and MaltParser only supports CoNLL-X, certain transformations were needed to perform the experiments. The corpus was flattened with conjoined tokens receiving a dummy surface form. The converted corpus is available alongside the original.[1]

To perform 10-fold cross validation we randomised the order of sentences in the corpus and split it into 10 equally-sized parts. In each iteration we held out one part for testing and used the rest for training. We calculated the labelled-attachment score (LAS) and unlabelled-attachment score (UAS) for each of the models.

The results we obtain are similar to those obtained with similar sized treebanks, for example the Irish treebank of Lynn et al. (2012), who report an LAS of 63.3 and a UAS of 73.3 with the best model.

[1] removedforreview

Features	**Algorithm**	**LAS** [range]	**UAS** [range]
surface	nivreeager	33.4 [23.4, 42.5]	50.9 [41.3, 63.2]
surface+lemma	nivreeager	33.4 [23.4, 42.5]	50.9 [41.3, 63.2]
surface+lemma+POS	nivreeager	55.2 [43.4, 75.0]	73.7 [65.8, 92.1]
surface+lemma+POS+MSD	nivreeager	55.2 [43.4, 75.0]	73.7 [65.8, 92.1]

Table 3: Preliminary parsing results from MaltParser using different models. The numbers in brackets denote the upper and lower bounds found during cross validation. Adding structural features to the model substantially improves the performance of the parser, although we find no improvement in using full morphosyntactic description (MSD) over using simply the first part-of-speech (POS) tag.

6 Future work

7 Conclusions

Acknowledgements

References

Atalay, Nart B., Kemal Oflazer, and Bilge Say (2003). “The Annotation Process in the Turkish Treebank”. In: *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora (LINC)*. Budapest, Hungary.

Ballesteros, Miguel and Joakim Nivre (2015). “MaltOptimizer: Fast and effective parser optimization”. In: *Natural Language Engineering* FirstView, pp. 1–27. ISSN: 1469-8110. DOI: 10.1017/S1351324914000035. URL: http://journals.cambridge.org/article_S1351324914000035.

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig, eds. (2015). *Ethnologue: Languages of the World*. Eighteenth. Dallas, Texas: SIL International.

Lynn, Teresa, Özlem Çentinoğlu, Jennifer Foster, Elaine Uí Dhonnchadha, Mark Dras, and Josef van Genabith (2012). “Irish Treebanking and Parsing: A Preliminary Evaluation”. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Istanbul, Turkey: European Language Resources Association (ELRA). ISBN: 978-2-9517408-7-7.

Marneffe, Marie-Catherine De, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning (2014). “Universal Stanford Dependencies: a Cross-Linguistic Typology”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Reykjavik, Iceland: European Language Resources Association (ELRA). ISBN: 978-2-9517408-8-4.

Nivre, J., J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi (2007). “MaltParser: A language-independent system for data-driven dependency parsing”. In: *Natural Language Engineering* 13.2, pp. 95–135.

Oflazer, Kemal, Bilge Say, DilekZeynep Hakkani-Tür, and Gökhan Tür (2003). “Building a Turkish Treebank”. English. In: *Treebanks*. Ed. by Anne Abeillé. Vol. 20. Text, Speech and Language Technology. Springer Netherlands, pp. 261–277. ISBN: 978-1-4020-1335-5. DOI: 10.1007/978-94-010-0201-1_15. URL: http://dx.doi.org/10.1007/978-94-010-0201-1_15.

Washington, Jonathan, Ilnar Salimzyanov, and Francis Tyers (2014). “Finite-State Morphological Transducers for Three Kypchak Languages”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn

Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Reykjavík, Iceland: European Language Resources Association (ELRA). ISBN: 978-2-9517408-8-4.

Балақаев, М., А. Исқақов, С. Кеңесбаев, Ғ. Мұсабаев, and Н. Сауранбаев (1954). *□азіргі □аза□ тілі : лексика, фонетика, грамматика*. Алматы: Қазақ ССР Ғылым Академиясы.