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

Lynn et al. (2012), Atalay et al. (2003), Oflazer et al. (2003), Marneffe et al. (2014)

2 Background

2.1 Kazakh

Kazakh (қазақ тілі), a Turkic language of Central Asia with around 12 million speakers, hargle bargle.

In the process of designing a Kazakh transducer (Washington et al., 2014), several decisions were made about how various inflectional morphemes affect the part of speech and syntactic function of various word forms. These decisions include the relations to other words defined by "case" morphemes, and 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. This means that indefinite accusative and genitive as well as nominative noun phrases are all unmarked for case, and hence are all ambiguous. Our 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 nominative noun phrases must have a corresponding predicate that agrees in person with the nominative noun phrase, e.g., a verb or a copula.

Copulas may also be non-overt. In the present tense, all but third person copulas have realised forms, but the third person copula does not surface.

Adjectives in Kazakh are by default attributive, but many may also be used either nominally (i.e., as nouns) or adverbially (i.e., as adverbs), or both. Our morphological transducer internally provides a series of classes;

depending on class membership, an adjective may receive 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.

2.2 Treebanks

3 Methodology

3.1 Corpus

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 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.

3.2.1 Tokens and words

4 Annotation guidelines

Label	Description

Table 2: Universal dependency label set

4.1 Copula

4.2 Coordination

4.3 Complex nominals

Some nominals occur together to act as a single nominal.

Features	LAS	UAS
-	-	-

Table 3: Preliminary parsing results from MaltParser using different models

One example of this is indefinite genitives. In such constructions, a "modifying" nominal is morphologically not marked (which means it will be tagged as `<nom>`, as discussed in §2.1), though should be considered an indefinite genitive. The modified element is a nominal with third-person possessive morphology, an example of which is provide in figure 1.

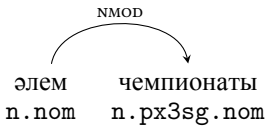

Figure 1: A nominal depending on an indefinite genitive.

In these examples, the indefinite genitive is considered `@nmod`.

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 and Nivre, 2015). The optimiser recommended the *covnonproj* parsing algorithm.

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 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 for unlabelled attachment 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`

6 Future work

7 Conclusions

Acknowledgements

References

Atalay, N. B., Oflazer, K., and Say, B. (2003). The annotation process in the Turkish treebank. In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora (LINC)*, Budapest, Hungary.

Ballesteros, M. and Nivre, J. (2015). MaltOptimizer: Fast and effective parser optimization. *Natural Language Engineering*, FirstView:1–27.

Lynn, T., Özlem Çentinoğlu, Foster, J., Dhonnchadha, E. U., Dras, M., and van Genabith, J. (2012). Irish treebanking and parsing: A preliminary evaluation. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Marneffe, M.-C. D., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal stanford dependencies: a cross-linguistic typology. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Oflazer, K., Say, B., Hakkani-Tür, D., and Tür, G. (2003). Building a Turkish treebank. In Abeillé, A., editor, *Treebanks*, volume 20 of *Text, Speech and Language Technology*, pages 261–277. Springer Netherlands.

Washington, J., Salimzyanov, I., and Tyers, F. (2014). Finite-state morphological transducers for three Kypchak languages. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavík, Iceland. European Language Resources Association (ELRA).