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[1]

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 402 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. In the article we present some issues in dependency syntax for Kazakh and how these are analysed in the universal-dependency framework. Preliminary results for statistical dependency parsing of Kazakh are reported, along with some directions for future research.

Keywords: Kazakh; treebank; dependency grammar; universal dependency

1 Introduction

This article describes work towards the development of a dependency treebank for Kazakh, a Turkic language spoken in Central Asia and Europe. 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 (Haverinen et al., 2013; Voutilainen, 2011) and Polish (Woliński et al., 2011) and smaller languages like Irish (Lynn et al., 2012). To our knowledge, however, a treebank exists for only one other Turkic 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 (De Marneffe et al., 2014).

The remainder of the paper is organised as follows. Section 2 gives some background linguistic information on Kazakh, and outlines some special challenges in parsing Kazakh. In Section 3 we describe the corpus that we annotated and the methodology used in annotating it. Section 4 gives a sketch of some decisions we have made with respect to annotation guidelines, referring back to the discussion in Section 2. 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 5, and in Sections 6 and 7 we give perspectives for future work and some concluding remarks.

[1]Corresponding author. E-mail address: jonwashi@indiana.edu

2 Background

2.1 Kazakh

Kazakh (қазақ тілі), a Turkic language of Central Asia and Europe, 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 overviews of the language, there is little to no work on the syntax of Kazakh within modern theoretical syntactic frameworks. The authors are familiar with such work on related languages, especially Turkish (e.g., Kornfilt, 1997 and Göksel and Kerslake, 2005); while not directly consulted for this work, these works have contributed to our understanding of Kazakh syntax.

As an agglutinative language with rich morphology and agreement phenomena, Kazakh presents some interesting challenges for computational syntax. These challenges include the syntactic functions of the various "case" morphemes, problems of "zero derivation", non-finite clauses, and copulas and copula constructions. An existing morphological transducer of Kazakh (Washington et al., 2014) implements analyses of how these various phenomena occur on the morphological level. These phenomena will be described in this section, and how they were dealt with in the annotation of the treebank will be described in section 4.

In Kazakh, as in most languages with case, there is not a one-to-one relation between "case" morphemes and syntactic function (not to mention a wide range of semantic functions). The main syntactic functions of the traditionally defined cases in Kazakh are summarised in table 1.

Table 1: Primary syntactic functions of traditionally defined cases in Kazakh.

Case	Morph	Functions	Examples
nom.	-	subject attributive indefinite object indefinite genitival	*Дәрігер үйді көреді.* 'The doctor sees the house.' *қонақ үй* 'hotel (lit., guest house)' *Дәрігер үй көреді.* 'The doctor sees a house.' *үй жануарлары* 'house animals'
acc.	-/NI/	definite object	*Дәрігер үйді көреді.* 'The doctor sees the house.'
gen.	-/NIң/	definite genitival embedded subject	*үйдің жануарлары* 'the animals of the house' *Ол Айгүлдің ойнағанына қарап тұр.* 'She's watching Aygül play.'
loc.	-/DA/	adverbial	*Үйде ұйықтап жатыр.* 'S/he's sleeping in the house.'
abl.	-/DAн/	adverbial comparator	*Дәрігер үйден шықты.* 'The doctor came out of the house.' *түйеден үлкен* 'bigger than a camel'
dat.	-/GA/	indirect object adverbial trans. causative subj.	*Ініме кітап бердім.* 'I gave my younger brother a book.' *Тойға көп кісі келіпті.* 'A lot of people came to the feast.' *Балаға әнді тыңдаттым.* 'I made the child listen to the song.'
inst.	-/Мен/	adverbial	*Олар күшікпен ойнап тұр.* 'They're playing with the dog.'

As seen in the table, the morphologically unmarked nominative case (e.g., *үй* 'the/a house' NOM) has a wide variety of uses, including indefinite object and genitival. Definite objects are marked with the accusative case (e.g., *үйді* 'the house' ACC), and definite genitivals are marked with the genitive case (e.g., *үйдің* 'of the house' GEN). The Kazakh transducer marks all bare nominals both as ‹NOM› and ‹ATTR›, but the various functions may be disambiguated syntactically. Genitival nominals, whether definite or not, must have a corresponding nominal with possessive morphology that agrees in person and number (e.g., *үйдің есігі* 'the/a door to the house'; *үй тапсырмасы* 'homework'), and subject nominals must have a corresponding predicate containing e.g., a verb or a copula that agrees in person and number with the nominative nominal (e.g., *үй көшіріледі* 'the house gets moved'). When a nominative nominal depends on another nominal that does not have possessive case, the first one must be considered attributive (e.g., *үй киім* 'house clothes').

The attributive use of nominals is similar to the use of adjectives, and could even be thought of as a "zero-

derivation" of nominals into adjectives. Interestingly, many adjectives can also be used substantively, i.e., as nominals. Table 2 shows the various functions that nominals, adjectives, and adverbs may take.

Table 2: The default (first line of each) and "zero-derived" uses of nominals, adjectives, and adverbs

Category	Function	Example
Nominals	Substantive	***Қонақтарымыз*** *бай.* '**Our guests** are rich.'
	Attributive	***қонақ*** *үй* 'hotel (lit., **guest** house)'
	Adverbial	*Оны бір* ***рет*** *көрдім.* 'I saw him/her/it one **time**.'
Adjectives	Attributive	***жақсы*** *үй* '**nice** house'
	Substantive	*ең* ***жақсы****лары* '**the nice**st **one**s'
	Adverbial	*Оны* ***жақсы*** *танимын.* 'I know him/her **well**.'
Adverbs	Adverbial	*Ол* ***кеше*** *кетті.* 'S/he left **yesterday**.'

Nominals (‹N›) default to substantive and adjectives (‹ADJ›) default to attributive. For the other readings, the transducer provides readings such as ‹ADJ›‹ADVL› and ‹N›‹ATTR›.

Kazakh, like most Turkic languages, makes frequent use of non-finite verb forms by deriving verbal adjectives, verbal nouns, and verbal adverbs from verbs. Verbal adjective phrases modify nouns, as in ***мені көрген*** *дәрігер* 'the doctor **that saw me**'. Verbal nouns can function as subjects or complements of verbs or copulas, as in *Дәрігер сені* ***көргенін*** *білген жоқпын.* 'I didn't know that the doctor had seen you.' and *Дәрігер сені* ***көргені*** *жақсы болыпты.* 'It's good that doctor saw you.' Verbal adverbs allow a verb phrase to function as a clausal verbal adjunct, as in *Мен* ***дәрігерді көріп*** *қуанып кеттім.* '**Seeing the doctor**, I got happy.' Verbal nouns with certain case morphology and/or occurring with certain postpositions can also function as a verbal adjunct much in the same way that verbal adverb clauses do, as in *Мен* ***дәрігерді көргенде*** *қуанып кеттім.* 'I got happy **when I saw the doctor**.', here with a verbal noun in the locative case.

Another category that may be non-overt is the copula. The primary strategy for copula constructions in Kazakh is the use of a defective verb *е-*. In the present tense, the verb itself does not surface, but agreement morphology surfaces (cliticised to the previous word) in all but the third person forms (e.g., *Мен үйде****мін****.* 'I'm at home.' versus *Ол үйде.* 'S/he's at home.'). The defective copula verb also surfaces in the recent past tense (e.g., *Мен үйде едім.* 'I was at home.', *Інім үйде еді.* 'My younger brother was at home.'). Of particular interest are non-finite forms of the copula. Copula clauses can be attributive, as in ***үйі жақсы*** *дәрігер* 'the doctor **who has a nice house**', or literally '**his/her house is nice** the doctor'. While this construction never has overt copula marking, copula clauses which are complements of verbs always have an overt copula form, e.g. *Дәрігердің үйі жақсы* ***екенін*** *білген жоқпын.* 'I didn't know **that** the doctor's house **was** nice.' Here, *екен* is a suppletive verbal noun form of the copula and is marked as accusative.

How each of these issues bears on a dependency analysis will be discussed in section 4.

2.2 Treebanks

A treebank is a parsed corpus of sentences annotated syntactically following a particular syntactic theory. Two broad groups can be distinguished: phrase-structure treebanks which annotate constituency strucure, and dependency treebanks which annotate dependency structure. Some treebanks combine both.

Treebanks can be used directly for linguistic and computational linguistic research by performing search queries—for example, to extract a valency lexicon for verbs, or to study the frequency of various syntactic phenomena such as word order or nominal case usage and syntactic function.

They can also be used to train statistical parsers which can be used to annotate previously unseen texts. These parsers can be used in end-user applications such as machine translation and computer-aided language learning. According to Nivre (2008), a parser trained on a treebank of only 1,500 sentences can provide reasonable parsing accuracy.

3 Methodology

3.1 Corpus

To create the corpus, we selected a range of texts from free and public-domain sources. These texts encompassed texts from a variety of genres, such as encyclopaedic articles, folk tales, legal texts, and phrases from a phrasebook. The corpus is not entirely balanced and leans more towards encyclopaedic text. Table 3 provides a breakdown of the sources.

Table 3: Composition of the corpus. The corpus covers a range of genres and text types from free & public-domain sources.

Document	Description	Sentences	Tokens	Avg. length
UN Declaration on Human Rights	Legal text on human rights	26	417	16.0
Phrasebook	Phrases from Wikitravel	38	204	5.4
Жиырма Бесінші Сөз	Philosophical text	33	467	14.2
Қожанасырдың тойға баруы	Folk tale from Wikisource	9	139	15.4
Ер Төстік	Folk tale from Wikisource	25	209	8.4
Азамат қайда?	Story for language learners	49	433	8.8
Футболдан әлем чемпионаты 2014	Wikipedia article (2014 World Cup)	12	191	15.9
Иран	Wikipedia article (Iran)	121	1714	14.2
Радиан	Wikipedia article (Radian)	2	17	8.5
Шымкент	Wikipedia article (Shymkent)	13	160	12.3
Wikipedia misc.	Random sentences from Wikipedia	74	565	7.6
		402	4516	11.2

Table 4 briefly details the various tags (i.e., syntactic functions) associated with the traditionally defined morphological cases of Kazakh as found in the corpus. Some mappings may be surprising, but will not be discussed in detail in this paper. It should also be noted that there are several known uses which are currently unattested (partitive use of ABL as DOBJ, embedded SUBJ marked as ACC) or rare (DAT as IOBJ) in the corpus.

Table 4: The functions associated with each case in Kazakh as currently found in the corpus.

	IOBJ	CMPND	NMOD	NMOD:POSS	DOBJ	ROOT	SUBJ	VOC
NOM	—	1	199	157	72	26	279	1
ACC	—	—	—	—	102	—	—	—
GEN	—	—	120	120	—	—	12	—
DAT	1	—	113	—	13	—	—	—
LOC	—	—	113	—	—	6	—	—
ABL	—	—	64	—	—	—	—	—
INS	—	—	33	—	—	—	—	—

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

Tokenisation of the corpus is performed by our morphological analyser. This analyser performs tokenisation on the basis of a left-to-right longest match algorithm described in Garrido-Alenda et al. (2002). Simple tokens such as *толқулар* 'riots' are maintained as a single token, and their lemma and morphological analysis is returned. Multiword units such as *ауа райы* 'weather' and *ата-анасының* 'of their parents' are combined into a single token. Abbreviations and numerals which bear case, such as *АҚШ-пен* 'with the USA' and *90%-ына* 'to 90%' are analysed as a single token, as are light verb constructions such as *пайда бол* 'to appear' and tense forms written with space like *оқыған жоқ*, the third-person negative past of *оқы-* 'to study/read'.

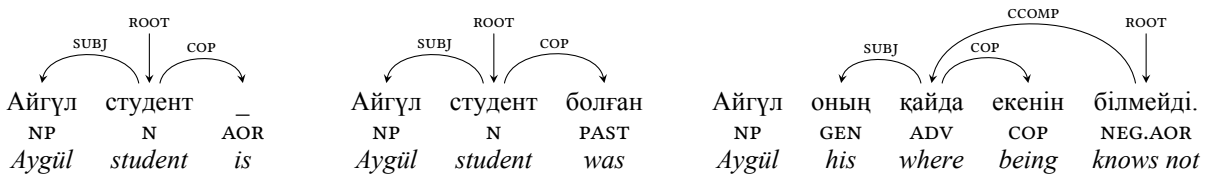
In some cases a single token is split into two tokens, as with the aorist copula suffixes, e.g., *үйдемін* 'I am in the house' is tokenised as *үй*.LOC + *е*.COP.AOR.SG1. Furthermore, two input tokens may result in three output tokens, e.g., *бар ма?* 'is there?' is tokenised as *бар*.ADJ + *е*.COP.AOR.SG3 + *ма*.QST.

4 Annotation guidelines

4.1 Copula

The copula (both *е-* and *бол-*) is a challenging problem for dependency analysis of Kazakh. The universal dependency guidelines state that the copula should be the dependent of the lexical predicate. However, in many cases the copula in Kazakh is found in embedded clauses, which morphologically acts much more like the head of the embedded clause.

We have uniformly annotated the copula as a leaf node with the predicate, or adverbial as the head of the structure. For certain structures this is convenient, such as the bare copula in phrases like 1a or 1b, but for phrases where the copula is part of an embedded clause this is not necessarily the most effective choice. In 1c, the copula holds all the morphological information, including agreement with the subject and accusative marking for the embedded clause.



(a) Non-surfaced copula. **(b)** Copula in past tense. **(c)** Copula in an embedded clause.

Figure 1: Dependency trees of copula constructions.

4.2 Coordination

One difference in our annotation scheme compared to the standard universal dependency analysis is with coordination. While the universal dependency scheme takes the first conjunct as the head, we take the last. This decision was made based on the fact that Kazakh is a head-final language and morphological marking is only obligatory on the last conjunct in a series. Furthermore, experiments in representing coordination in other predominantly head-final languages have found that the final-conjunct head analysis results in better parser accuracy (Bengoetxea and Gojenola, 2009). Figure 2 shows an example of our coordination strategy.

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 (morphologically 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

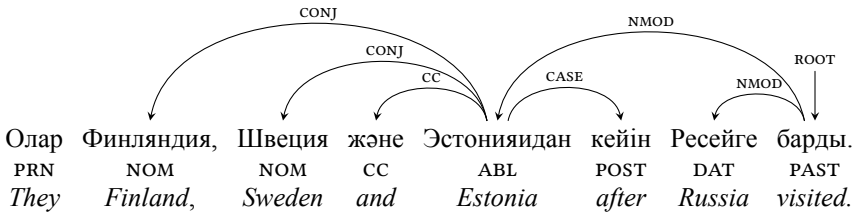
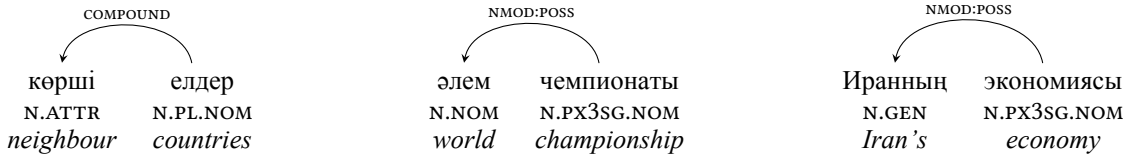


Figure 2: Coordination: All conjucts are attached to the final conjunct, which is the head of the coordinated phrase.

indefinite genitive nominal (morphologically 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 genitive construction is formed by a genitive-marked nominal (tagged ‹GEN›) precding a nominal that has third-person possessive morphology, as shown in 3c.



(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:POSS, regardless of whether there is a definite or indefinite genitive construction.

4.4 Non-finite clauses

As discussed in section 2.1, Kazakh makes extensive use of non-finite clauses, including verbal adjective clauses, verbal noun clauses, and verbal adverb clauses.

Verbal adjective clauses modify a head nominal, effectively allowing a whole verb phrase to act as an adjective. The dependency relation between them is ACL, per UD documentation. An example is provided in figure 4.

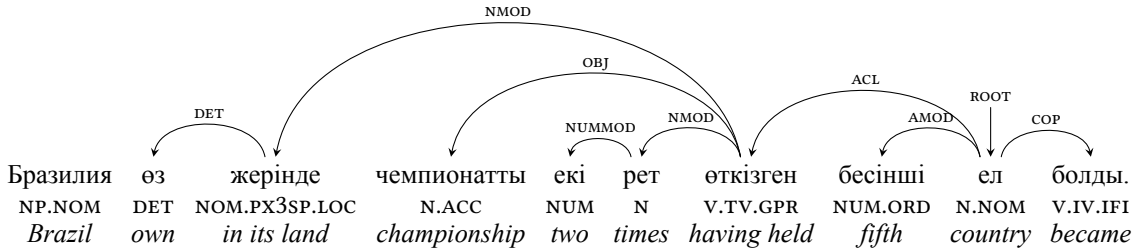


Figure 4: Verbal adjectives: verbal adjectives are the head of everything in their own clause and are an ACL dependency of the noun they modify.

Verbal adverb clauses in Kazakh act as a clausal verbal adjunct, essentially allowing a whole verb phrase to act as an adverb. The dependency relation between the verbal adverb and the head verb is ADVCL, per UD documentation. An example is provided in figure 5.

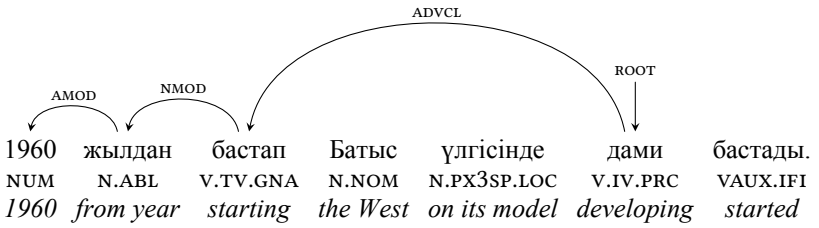


Figure 5: Verbal adverbs: verbal adverbs are the head of everything in their own clause and are an ADVCL dependency of the verb they are subordinate to.

Verbal nouns allow a whole verb phrase to be used as a nominal, and the resulting verbal noun phrase may be used as a subject or object of a verb. When used as the subject of a verb, verbal noun phrases receive the UD label SUBJ. When used as the object of a verb, verbal noun phrases receive the UD label CCOMP, unless the subject of the subordinate clause is obligatorily identical to the subject of the main clause, in which case it receives XCOMP. An example of a verbal noun phrase used as a CCOMP dependency of the main verb is given in figure 6.

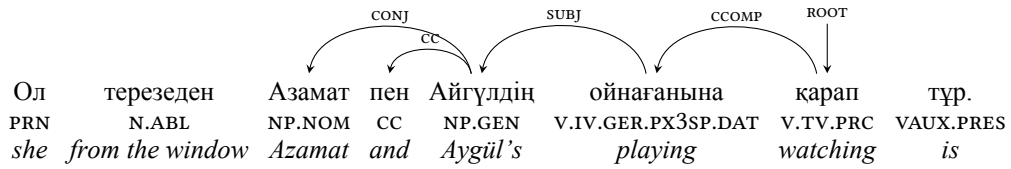


Figure 6: Verbal nouns: verbal nouns are the head of everything in their own clause and can be a CCOMP dependency of the verb they are the complement of. Note here that the subject noun phrase of the subordinate clause is in genitive case.

As mentioned earlier, verbal nouns combined with certain case morphology and/or postpositions can function as an adverbial adjunct to a main clause, similar to verbal adverb clauses. In such instances, they receive the UD label ADVCL, and example of which is given in figure 7.

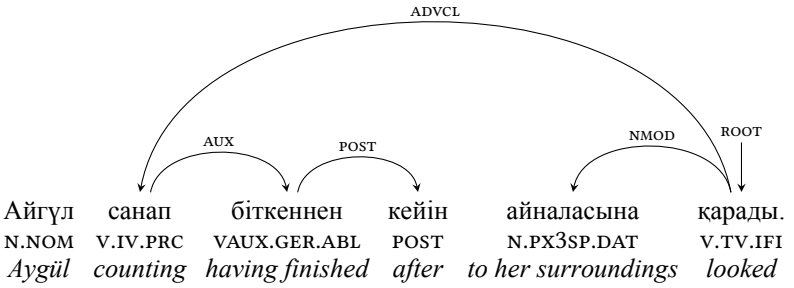


Figure 7: Verbal nouns with case/postpositions: verbal nouns with certain case morphology and/or postpositions are the head of everything in their own clause and can be an ADVCL dependency of the verb they are the complement of.

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 was run separately for each of the model configurations, and the best combination of parameters was selected.

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.[2]

Table 5: Preliminary parsing results from MaltParser using different models. The numbers in brackets denote the upper and lower bounds found during cross-validation.

Features	**Algorithm**	**LAS** [range]	**UAS** [range]
surface	nivreeager	42.8 [38.9, 45.9]	58.4 [52.4, 62.9]
surface+lemma	nivreeager	42.8 [38.9, 45.9]	58.4 [52.4, 62.9]
surface+lemma+POS	nivreeager	64.9 [66.4, 67.1]	77.0 [76.8, 80.0]
surface+lemma+POS+MSD	nivreeager	76.8 [70.9, 80.0]	81.4 [75.7, 85.2]

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, shown in table 5, are similar to those obtained with similar sized treebanks, for example the Irish treebank of Lynn et al. (2012), for which an LAS of 63.3 and a UAS of 73.3 were reported with the best model. Adding structural features to the model substantially improves the performance of the parser.

6 Future work

Future work will focus on improving the annotation guidelines and the consistency of annotation in the corpus. We will also study the possibility of deepening the annotation with Turkic-specific relations. When we have stable annotation guidelines we intend to extend the corpus with more texts. We would also like to work on cross-lingual dependency parsing, that is, applying a model learnt on the Kazakh treebank to other Turkic languages such as Tatar, Kumyk, and Tuvan. We have morphological analysers for these languages which have compatible tagsets for morphological features and as such it should be possible to learn a delexicalised model based on these features. As Turkic syntax is broadly homogenous, this presents a promising avenue for future work.

7 Concluding remarks

We have presented the first steps towards a free/open-source dependency treebank for Kazakh with annotation based on the universal dependencies. The treebank is small, but provides a base for bootstrapping further. Performance of a state-of-the-art statistical parser trained on the treebank is comparable to other treebanks of similar size.

Acknowledgements

The authors would like to acknowledge Tolgonay Kubatova; Zhenisbek Assylbekov, Aida Sundetova, and colleagues; and Aibek Makazhanov for the various ways in which they each contributed to this research.

[2] http://svn.code.sf.net/p/apertium/svn/languages/apertium-kaz/texts/puupankki/

References

Ballesteros, Miguel and Joakim Nivre (2015). "MaltOptimizer: Fast and effective parser optimization". In: *Natural Language Engineering* FirstView, pp. 1–27.

Bengoetxea, Kepa and Koldo Gojenola (2009). "Exploring Treebank Transformations in Dependency Parsing". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pp. 33–38.

De Marneffe, Marie-Catherine, 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 9th International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by N. Calzolari, Kh. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis. Reykjavik, Iceland.

Garrido-Alenda, Alicia, Mikel L. Forcada, and Rafael C. Carrasco (2002). "Incremental construction and maintenance of morphological analysers based on augmented letter transducers". In: *Proceedings of the Conference on heoretical and Methodological Issues in Machine Translation*, pp. 53–62.

Göksel, Aslı and Celia Kerslake (2005). *Turkish: A Comprehensive Grammar*. London: Routledge.

Haverinen, Katri, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter (2013). "Building the essential resources for Finnish: the Turku Dependency Treebank". In: *Language Resources and Evaluation*. In press. Available online., pp. 1–39.

Kornfilt, Jaklin (1997). *Turkish*. London: Routledge.

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 8th International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by N. Calzolari, Kh. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis.

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.

Nivre, Joakim (2008). "Algorithms for deterministic incremental dependency parsing". In: *Computational Linguistics* 34, pp. 513–553.

Oflazer, Kemal, Bilge Say, Hakkani-Tür, Dilek Zeynep, 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.

Voutilainen, Atro (2011). "FinnTreeBank: Creating a research resource and service for language researchers with Constraint Grammar". In: *Proceedings of the NODALIDA 2011 workshop Constraint Grammar Applications*.

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 N. Calzolari, Kh. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis. Reykjavík, Iceland.

Woliński, Marcin, Katarzyna Głowińska, and Marek Świdziński (2011). "A Preliminary Version of Składnica—a Treebank of Polish". In: *Proceedings of the 5th Language & Technology Conference*. Ed. by Zygmunt Vetulani. Poznań, Poland, pp. 299–303.

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