A Dependency Treebank for Buryat

Elena Badmaeva
Informatika fakultatea
Euskal Herriko Unibertsitatea
Donostia
lenab23.02@gmail.com

Francis M. Tyers
Giela ja kultuvrra instituhtta
UiT Norgga árktalaš universitehta
Romsa
francis.tyers@uit.no

Abstract

There has been little research on natural-language processing of Buryat, in part due to the absence of language resources. In this article we present the first syntactic treebank for Buryat. It is an evolving project based on the Universal Dependencies (UD) annotation guidelines. We report on the procedure of constructing the treebank and explain some features of the labelling scheme with respect to linguistic phenomena in Buryat. The annotation of 919 sentences was done manually from scratch without using any automatic labelling tools. We evaluate the performance of various language models with UDPipe and describe the plans for future work.

1 Introduction

In this article we present the first publicly available treebank for Buryat, the second official and national language of the Republic of Buryatia, located within Russia in Southern Siberia. According to UNESCO report, Buryat is considered to be an endangered language and at risk to disappearing (Skribnik, 2006). Even though Buryat is reported to be one of the most investigated Mongolic languages (Skribnik, 2006), comparably little computational linguistic research has been maintained for it. To our knowledge, among the Mongolic languages only Inner Mongolian has a treebank (Loglo et al., 2014).

Although there are no computational tools, Buryat possesses an extensive range of written-language data, it is the only language in Siberia which has its own historical records and there are regularly published Buryat newspapers, journals, books, films, television and radio programmes. The development of language technologies for Buryat is an essential step towards its revitalisation and documentation (Badagarov et al., 2016). To this end, we report on our efforts to build the first Buryat dependency treebank following the guidelines of Universal Dependencies (UD), a cross-lingual treebank annotation project (Nivre et al., 2016). We chose the UD scheme for the annotation as it provides ready-made recommendations on which to base annotation guidelines. This reduces the amount of time needed to develop

bespoke annotation guidelines for a given language as where the existing *universal* guidelines are adequate they can be imported wholesale into the language-specific guidelines.

The remainder of the paper is organised as follows. Section 2 gives some background typological information on Buryat. Then section 3 gives an overview of the corpus used for the treebank and how the annotation was done. Section 4 discusses some features of Buryat and how they were dealt with in the annotation guidelines. Furthermore, in section 5 we report preliminary parsing experiments for Buryat using the state-of-the-art NLP pipeline UDPipe. We test different combinations of features. In section 6 we present a summary and suggest avenues for future research and development.

2 Buryat

Buryat (in Buryat *Буряад хэлэн*) is a Mongolic language spoken by the Buryats (Skribnik, 2006). The greater part of Buryat population live in the Buryat Republic which is located in the southern part of Siberia around Lake Baikal. According to the census of 2010, there are about 461,389 Buryat people in Russia (Росстат, 2010). Like the other languages of the Mongolic group Buryat is an agglutinative language with Subject-Object-Verb constituent order (Fuss, 2005; Skribnik, 2006). However unlike Khalkh Mongolian, Buryat has verbal agreement with the subject. Example 1 is a sentence in Buryat showing SOV word order features.

(1) Буряад зон бултадаа Байгал далайда дуратай.
Buryat people all Baikal lake love.
'All Buryat people love lake Baikal .'

There are five main dialect groups of Buryat, but the literary standard is based on the Khori dialect group, spoken to the east of Lake Baikal. This is the dialect group with the greatest number of speakers (Skribnik, 2006).

Buryat has seven cases, and two numbers. Nominative case and singular number are unmarked. There are a large number of non-finite verb forms (both participles and verbal adverbs) which are used for clause-level subordination and forming relative clauses.

3 Corpus

Currently the Buryat treebank consists of 919 annotated dependency trees and 10,146 tokens. The annotation was made using *Brat*, an online tool for text annotation (Stenetorp et al., 2012). The total amount of time spent on the creation of the treebank was nine months. This involved two people, one annotator and one supervisor. The process including the following steps: review of the relevant literature, discussion, consultation with experts on Buryat linguistics and contributors to the UD

project, annotation, correction, and lemmatisation. The annotation process itself was done by the first author over a period of three months.

Source	Domain	Sentences	Tokens	Avg. length
udtwenty	Grammar	20	152	7.6
wikipedia	Encyclopaedic	31	343	11
proverbs	Proverbs	3	23	7.6
buryad-unen	Newspaper	667	8,239	12.3
translation	Grammar	198	1,389	7
Total:		919	10,146	9.1

Table 1: Composition of texts in the Buryat Dependency Treebank. The *grammar* domain includes grammar-book style sentences.

Statistics about the corpus can be found in Table 1. The text source `udtwenty` consists of 20 sentences illustrating different grammatical features available through the UD project. The `wikipedia` text source consists of sentences extracted from the Buryat Wikipedia. Given their public domain status, we also included three Buryat proverbs from the `proverbs` source. Finally, after we were given permission to redistribute them under a free licence, we included articles from the newspaper *Buryaad Ünen* and some translations of example sentences from the *Technical Report Syntax Annotation Guidelines* for the Turku Dependency Treebank (Haverinen, 2012).

Having translated text as such a large portion of the treebank has three motivations. The first is that we aim to expand the treebank with more sentences from different domains of Buryat, so this percentage will decrease with time. The second is that we wanted to cover a wide range of syntactic structures, and the third is that we wanted to include translated text as a domain in itself.

3.1 Preprocessing and lemmatisation

Before annotating in the `brat` tool, the corpus was preprocessed and manually tokenised. Tokenisation was on space after splitting punctuation characters. The automatic tokenisation for each sentence was fixed manually before the sentence was annotated.

In order to add lemmas to the corpus, we created a lookup table of tuples (surface form, part-of-speech) to lemma. For example:

(харандаашууд, NOUN) → харандааш
(*pencils*, NOUN) → *pencil*

We then applied this lookup table to the corpus deterministically, for each (surface form, part-of-speech) pair inserting the lemma found in the lookup table. We

discovered no ambiguity in lemmatisation. The lookup table is available online under a free/open-source licence.[1]

4 Annotation guidelines

The annotation guidelines for Buryat were based on Universal Dependencies (Nivre et al., 2016). This is an international collaborative project to make cross-linguistically consistent treebanks available for a wide variety of languages. In the UD annotation scheme, to improve cross-linguistic compatibility, dependency relations are primarily between content words, with function words attaching as leaf nodes. The motivation for this is that content words are more stable between languages, while languages can vary with how e.g. cases are used as opposed to adpositions, and analytic versus synthetic tense constructions. Thus, in auxiliary–main verb constructions, the main verb is the head and the auxiliary is attached as a dependent, if there is more than one auxiliary they are attached as siblings as opposed to a nested structure. In adpositional phrases, the complement of the adposition is the head and the adposition itself is a dependent attached with the `case` relation.

In the following subsections we describe some examples from the Buryat treebank, making emphasis on the most relevant and more typologically interesting aspects. Thus, for reasons of space, we have omitted a discussion of finite-verbs in simple clauses as they function more or less as one may expect. Further information on the annotation guidelines can be found in Badmaeva, (2016).

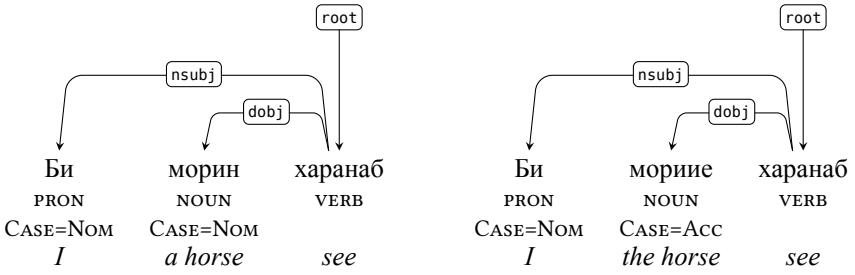
4.1 Core and oblique nominals

As previously mentioned, Buryat has seven cases: Nominative, accusative, genitive, dative,footnoteThe dative case in Buryat has double use as dative and locative. It is sometimes referred to as dative/locative, but in our work we refer to it simply as dative. ablative, instrumental and possessive. All of these apart from nominative are marked with suffix morphology.

The first two can be considered core cases. They have the following syntactic functions: The nominative case is used for the subject and the non-specific direct object of a simple clause. The overt accusative suffix is used for a specific direct object. Regardless of the case used the subject depends on the verb with the label `nsubj` and the direct object depends on the verb with the label `dobj` (see Figure 1).

The oblique cases are the dative, ablative, instrumental. The dative and ablative are locational cases used for both spatial and temporal meanings. In addition the dative is used in some non-verbal predicate constructions (see section 4.3) and as the case of the actor in passive constructions, and the ablative is used as the referent of comparison. In all cases these are marked with the `nmod` label, and when the locative is used to indicate possessor it is marked with `nmod:own`.

[1] `https://svn.code.sf.net/p/apertium/svn/incubator/apertium-bua/dev/lexicon.bxr.tsv`



(a) Non-specified direct object **(b)** Specified direct object

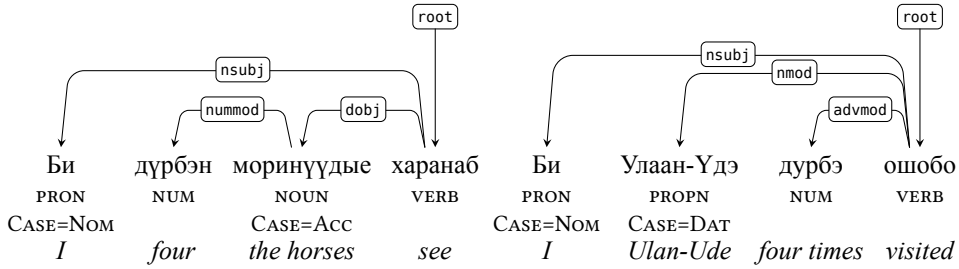
Figure 1: Subject and object marking in main clauses. The nominative is used for non-specified direct objects and the accusative is used for specified direct objects.

When formed from animate nouns the instrumental indicates the active participation of a second participant in the action, while when formed from inanimate nouns it can indicate use of a tool, means of transport, time period, etc. In both of these it receives the label `nmod`.

The genitive case expresses various kinds of adnominal attribution. Genitive modifiers precede their heads. Nouns in the genitive case depend on the nominal they modify and receive the label `nmod`. The possessive case may be used both adverbially, where it denotes accompaniment, and adnominally where it denotes simple possession. In both cases it receives the label `nmod`

4.2 Numeral expressions

Numerals have a regular case paradigm and can be used attributively, adverbially and predicatively. When used attributively to indicate quantity then they receive the label `nummod`, when used adverbially to indicate a number of times they receive the label `advmod` and when used predicatively in an equative construction they are the root of the sentence.



(a) Numeral expressing quantity **(b)** Adverbial numeral use

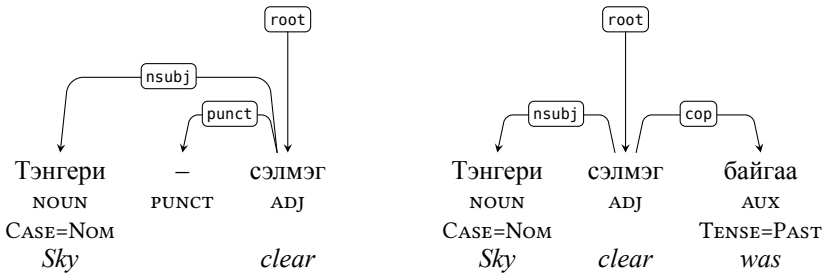
Figure 2: Numeral expressions in Buryat.

There are a number of other uses of numerals, and these are annotated according to their function.

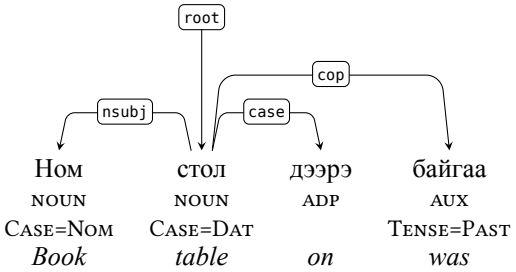
4.3 Non-verbal predication

In Buryat, constructions involving non-verbal predicates (nominals, adverbials) require a copula. This copula (a form of *бай-* 'to be') can be omitted in the present tense (see Figure 3). Constructions with a copula can be split into a number of categories: Equation, attribution, location, possession and existential.

Equation describes sentences such as *Борбилоо – шубуун* 'The sparrow is a bird' (lit. *Sparrow – bird*) where the predicate is a noun. Attributive sentences such as Figures 3a and 3b are where the predicate is an adjective. Locational sentences have a locative nominal or adverbial as the head (see Figure 3c). In all cases the non-verbal predicate is the head of the clause, and the copula is a dependent on the predicate with the label `cop`.



(a) Copula omission **(b)** Overt copula

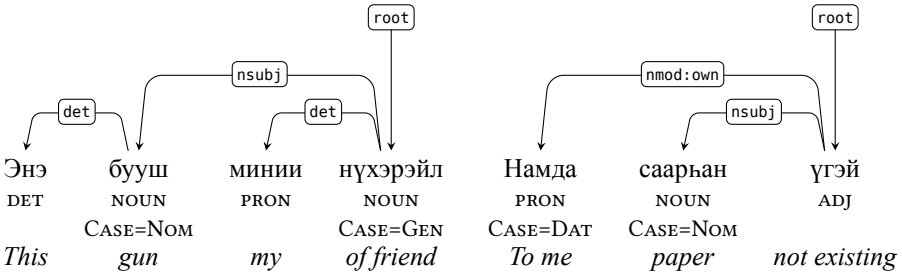


(c) Locative with copula

Figure 3: Equative, attributive and locative nominal predication

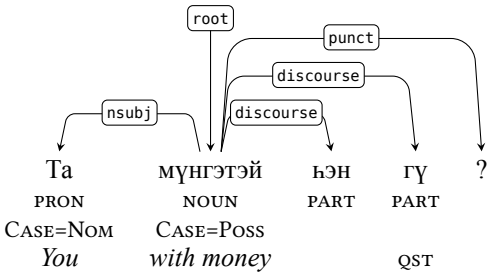
Existential sentences have a similar structure to locational sentences, but with reversed word order. In addition, in existential sentences the existential nominals *буй* 'existing', *бии* 'existing' (emphatic) or *үгэй* 'not existing' may be employed. These then become the head of the clause, as in (4b).

Possessive sentences do not have their own structure, and instead there are three patterns for possessives, two of which follow the attribution pattern (4a and 4c) and one which follows the locational pattern (4b).



(a) Predicative use of genitive

(b) Locative possessive



(c) Predicative use of possessive

Figure 4: Possible constructions for possessive in Buryat: (a) 'This gun belongs to my friend', (b) 'I do not have paper', and (c) 'Did you have money ?'.

In the attribution pattern and in bare locative possession (i.e. without *биш*, *буй* or *үгэй*), the item being owned is the head of the clause and the owner is the subject, while in the locative/existential construction, the item being owned is the subject and the owner in the dative case is annotated with the relation `nmod:own`.

4.4 Non-finite clauses

Buryat has a large number of morphemes for participles and verbal adverbs[2] which are marked for tense, aspect and mood.

Participle clauses modify a head nominal, effectively allowing a whole verb phrase to act as an adnominal modifier. This is the way in which Buryat forms relative clauses, so we annotate them with the dependency relation `acl`, per UD documentation. Examples are provided in Figures 6 and 5. There are also headless relative clauses where the inflection that would normally attached to the head nominal appear on the participle.

[2]In much of the Mongolic and typological literature verbal adverbs are referred to as *converbs*.

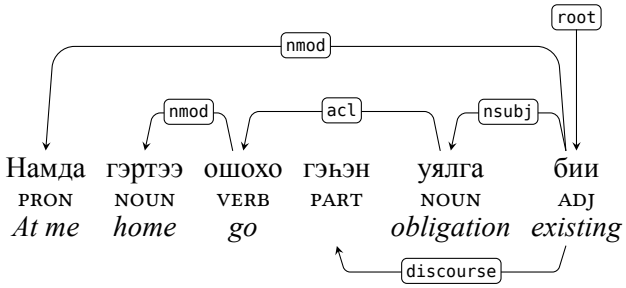


Figure 5: Relative clause with participle, 'I must go home.'

Verbal adverbs create adverbial subordinate clauses which may have an independent subject. These receive the label `advcl`. When they have an independent subject, this is marked with possessive suffix on the verb.

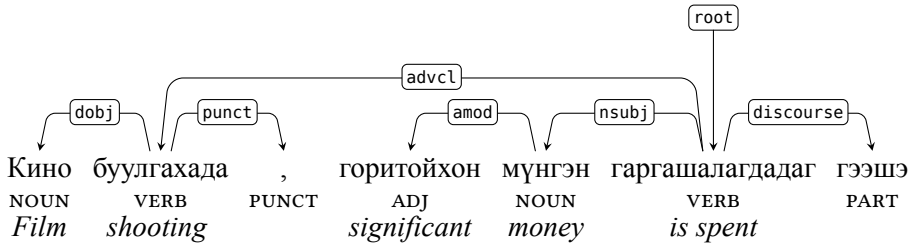


Figure 6: Adverbial clause with the verbal adverb suffix for 'when', 'When shooting a film, significant money is spent'.

4.5 Headedness

In contrast to the universal dependency rule about headedness in conjunction and multiword constructions, which prescribes taking the first token as the head, in Buryat we take the last token as the head. This is because in Buryat only the last word in the phrase is inflected. In line with treebanks for the Turkic languages, such as Kazakh (Tyers et al., 2015), we decided to annotate these as right-headed.

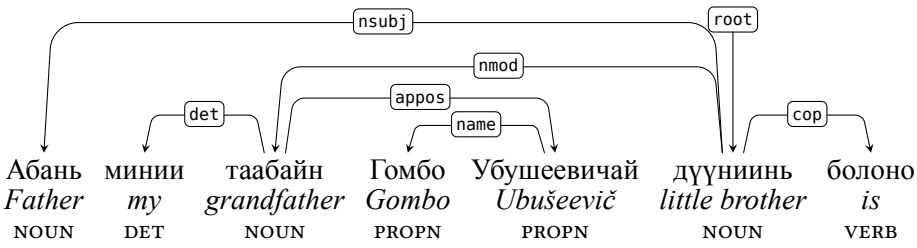


Figure 7: Dependency tree for the sentence "The father is a little brother of my granddad Gombo Ubušeevič." Note how the appositive proper noun sequence is right headed.

4.6 Discourse words

Buryat possesses a great deal of uninflected words, traditionally called *частицануудэ* 'particles' which may encode the following categories: negation, mood, tense (Skribnik, 2006). Predicates are often accompanied with the predicative particles that are subdivided into interrogative, negative, modal, evidential and copular particles. Some of the postpositional particles developed into clitics or suffixes. Though negation is usually expressed through suffixation (morpheme *-гүй*) in non-predicative sentences negative particles *үгы* and *бүү* are also used (Skribnik, 2006).

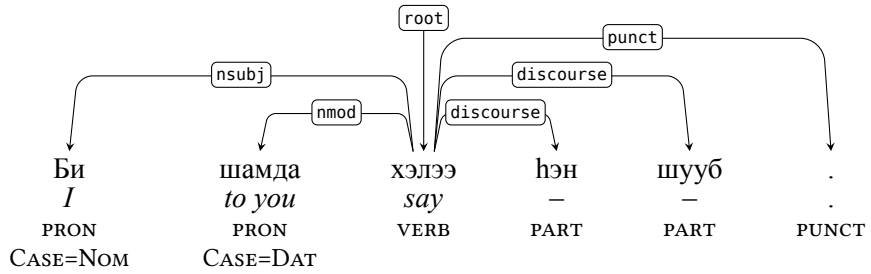


Figure 8: Dependency tree for the sentence "I did tell you." (lit. *I did say (it) to you*). The word *һэн* is a past marker, while *шууб* is an emphatic marker.

Following the Universal Dependencies guidelines we attach these to the head of the most relevant nearby clause in a flat structure.[3]

5 Experimental results

In order to test the treebank in a real setting, we evaluated a parser trained with UDPipe (Straka et al., 2016). UDPipe is an open-source trainable pipeline for tokenisation, tagging, lemmatisation and dependency parsing. It does not require any language-specific knowledge.

The options to UDPipe were that we trained a single model for parsing and two models for tagging. The reason for this is that according to the documentation, the system works better if a different models are used for predicting part-of-speech tags and lemmas.

We first held out 19 sentences at random in order to have a neatly divisible number of sentences. Then we performed 10-fold cross-validation by randomising the order of sentences in the corpus and splitting them into 10 equally-sized parts. In each iteration we held out one part for testing (90 sentences) and used the rest for training (810 sentences). We calculated the labelled-attachment score (LAS) and unlabelled-attachment score (UAS) for each of the models. In addition, for those

[3] We agree with one reviewer who suggested that annotating these with the `discourse` relation makes it difficult to separate interjections from grammatical function words. In version 2.0 of the treebank these will be annotate these with the relation `aux` in keeping with the new guidelines.

combinations of features that do not include the lemmas and part-of-speech tags from the gold tags we also calculate the part-of-speech tagging and lemmatisation accuracy.

To test how well models trained from the treebank perform, we made a number of configurations of the testing data. The first configuration was just to use the surface forms in the input, with the model providing lemmas and part-of-speech tags. This gives the current state-of-the art results in end-to-end processing of Buryat for both lemmatisation, part-of-speech tagging and dependency parsing.

The second configuration was to use the surface form and part-of-speech in the input. The objective of this configuration was to have a baseline that we could use to compare with the configuration where we include lemmas. The final configuration was to use both lemmas and part-of-speech from the treebank. This can be considered the upper-bound on parsing performance using the treebank.

Input features	**Lemma acc.**	**POS acc.**	**UAS**	**LAS**
Surface	[78.5, 85.9]	[76.0, 78.9]	[63.6, 69.5]	[45.1, 54.3]
Surface+POS	–	–	[69.2, 75.4]	[57.2, 63.7]
Surface+Lemma+POS	–	–	[70.7, 76.3]	[59.2, 64.3]

Table 2: Preliminary parsing results obtained with UDPipe for a range of metrics

The results of the experiments are presented in Table 2. While the results for end-to-end parsing are far from the state of the art, they provide a useful baseline for future work on natural language processing of Buryat.

We are unfortunately not able to compare performance on tools trained on our treebank with the Mongolian treebank of Loglo et al., (2014) as their treebank is not freely available, although the same authors in Loglo et al., (2013) report an accuracy of 75.21% UAS for rule-based dependency parsing.

6 Concluding remarks

In our paper we have presented the first dependency treebank for Buryat, and the first publically available treebank for a Mongolic language. It is the fundamental step towards the further development of natural-language processing and language technology for Buryat. During the process a number of discussions about some linguistic phenomena in Buryat were initiated which will be further investigated.

There are a number of avenues for future work. The first is to include morphological features. Buryat is a highly inflecting language and it has been shown for highly-inflecting languages that including morphological features makes it possible to learn more accurate parsing models. To this end we intend to build a finite-state transducer compatible with the treebank to do morphological analysis. Secondly, following the release of the version 2.0 guidelines for Universal Dependencies, we plan to convert the treebank to the new guidelines. Many aspects will be able to be

automatically converted, but we expect to manually revise all cases of ellipsis. Finally, we intend to synthesise the contents of this paper and MA thesis (Badmaeva, 2016) into a set of comprehensive online guidelines for Buryat.

Acknowledgements

The authors would like to express their gratitude to the editors of the newspaper *Buryat Unen* who gave us permission to include text from their articles for our treebank. We would like to thank Zhargal Badagarov and Dizhyd Markhadaeva who walked us through the most complicated issues of Buryat linguistics. We thank Jonathan Washington for his recommendation of the grammar sources on Mongolic languages and Martin Popel for his help to adjust UDPipe for our language model. We are also thankful to Koldo Gojenola and Gosse Bouma for sharing with us their valuable views on our project. In addition we are grateful to the anonymous reviewers whose detailed comments and suggestions allowed us to improve this article.

References

Badagarov, J., Trosterud, T., and Tyers, F. (2016). Language Documentation and Language Technologies for Circumpolar Region. URL: `http://www.uarctic.org/media/1176610/badagarovthematic-network-language15june2015.pdf` (visited on 08/09/2016).

Badmaeva, E. (2016). Universal Dependencies for Buryat. MA thesis. The University of the Basque Country, Spain.

Fuss, E. (2005). The Rise of Agreement: A Formal Approach to the Syntax and Grammaticalization of Verbal Inflection. John Benjamins Publishing.

Haverinen, K. (2012). Syntax Annotation Guidelines for the Turku Dependency Treebank. University of Turku, Department of Information Technology.

Loglo, S. and Sarula (2013). A Rule-Based Mongolian Dependency Parsing Model. In: *International Journal of Knowledge and Language Processing* 4.3, pp. 27–37.

— (2014). Construction of a Mongolian Dependency Treebank. In: *International Journal of Knowledge and Language Processing* 5.2, pp. 32–42.

Nivre, J. et al. (2016). Universal Dependencies 1.3. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague. URL: `http://hdl.handle.net/11234/1-1699`.

Skribnik, E. (2006). Buryat. In: *The Mongolic Languages*. Ed. by J. Janhunen. Routledge, pp. 102–127.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. In: *Proceedings of the Demonstrations Session at EACL 2012*. Avignon, France: Association for Computational Linguistics.

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA). ISBN: 978-2-9517408-9-1.

Tyers, F. and Washington, J. (2015). Towards a Free/Open-Source Universal-Dependency Treebank for Kazakh. In: *Proceedings of the International Conference "Turkic Languages Processing" (Turklang 2015)*.

Росстат (2010). Russian Census.