Proceedings of SIGUL2022 @LREC2022, pages 185–188
Marseille, 24-25 June 2022

Introducing YakuToolkit

Yakut Treebank and Morphological Analyzer

Tatiana Merzhevich, Fabrício Ferraz Gerardi
Universität Tübingen
Seminar für Sprachwissenschaft
{tatiana.merzhevich, fabricio.gerardi}@uni-tuebingen.de

Abstract

This poster presents the first publicly available treebank of Yakut, a Turkic language spoken in Russia, and a morphological analyzer for this language. The treebank was annotated following the Universal Dependencies (UD) framework and the morphological analyzer can directly access and use its data. Yakut is an under-represented language whose prominence can be raised by making reliably annotated data and NLP tools that could process it freely accessible. The publication of both the treebank and the analyzer serves this purpose with the prospect of evolving into a benchmark for the development of NLP online tools for other languages of the Turkic family in the future.

Keywords: Yakut, Sakha, Turkic languages, Universal Dependencies, Morphology, NLP, Finite State Morphology

1. Introduction

Yakut or Sakha (ISO sah, Glottocode yaku1245) is the easternmost member of the Turkic language family, spoken in the Republic of Sakha (Yakutia) in the Far Eastern Federal District of Russia. The distribution of Turkic languages, taken from Glottolog 4.5 (Nordhoff and Hammarström, 2011) is shown in Figure 1 with Republic of Sakha colored in green. In spite of their broad geographical distribution, all Turkic languages including Yakut are head final languages sharing features like SOV word order, agglutinative morphology, synthetic structure, and syllabic harmony. Although Yakut is not intelligible to speakers of other Turkic languages. Nonetheless, all Turkic languages still share many structural features that clearly allow then to be identified as Turkic (Johanson, 2021; Menz and Monastyrev, 2022).

The Federal State Statistics Service[1] estimated the population of the Republic of Sakha to be about 1 million people in 2021. Of these, the half is considered to be native Yakuts. Based on the 2002 census (Eberhard et al., 2021), 93% of the ethnic population speak Yakut and the language enjoys the official status of a provincial language and is thus used in education, work, mass media, and administration (Eberhard et al., 2021). Nonetheless, at the same time it is also categorized as an endangered language (ELP, 2020; Moseley, 2010), partly due to the increasing use of Russian among younger generations. The gradual loss of Yakut speakers can be indirectly seen in the higher density of monolingual speakers in rural areas.

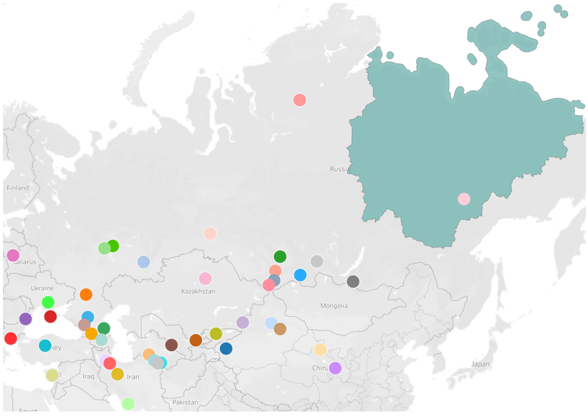


Figure 1: Distribution of Turkic languages according to Glottolog 4.5. Each language is represented by a single dot and a unique color. Yakut is spoken in the green shaded area.

Within the Turkic family, the importance of Yakut is evident due to its being the only language, besides Turkmen and Khalaj, to have maintained traces of primary vowel-length distinction (Johanson, 2021); and the presence of borrowings from Mongolic and Russian, with a Tungusic and Yeniseic substratum (Menz and Monastyrev, 2022). Still, although Yakut is used in education and public life, it can be considered to be an under-represented language. The major linguistic descriptions of the language are mainly available in Russian and, to our knowledge, little to no online NLP tools are able to process Yakut.

The lack of open access tools was the primary motivation behind the work on the Universal De-

[1] https://rosstat.gov.ru. Accessed on 16/04/2022.

pendencies Yakut treebank. By making syntactically and morphologically annotated texts of different genres and complexity available, the treebank will allow for more comprehensive understanding of both Turkic languages and languages in general. At the same time it serves as a departing point for the creation of NLP tools, which are practically non-existent. Parallel to the Yakut treebank we are also working a finite-state morphological analyzer which extends the potential of NLP tasks that can be carried out for Yakut.

Among available tools for Yakut we are aware of the following: 1) the morphological analyzer and generator for Sakha (WiN, 2021), 2) annotated morphological data, which is a part of the Universal Morphology project (Kirov et al., 2018), 3) an online Sakha-Russian-Sakha dictionary, which is apparently being expanded with English translation (Anonymous, 2012).

The rest of this paper is organized as follows: Section 2 introduces the UD-Yakut treebank, and Section 3 introduces the morphological analyzer. Section 4 concludes the papers with some brief remarks.

2. The UD-Yakut Treebank

Universal Dependencies (De Marneffe et al., 2021) is a multilingual formalism which offers annotation guidelines[2] for dependency relations, morphological analysis, part-of-speech tagging, among others. Despite some drawbacks of UD (Osborne and Gerdes, 2019), it is arguably the best open-access framework available nowadays. Alternatives such as SUD (Gerdes et al., 2018) are also worth considering and a conversion and parallel maintenance is planned.

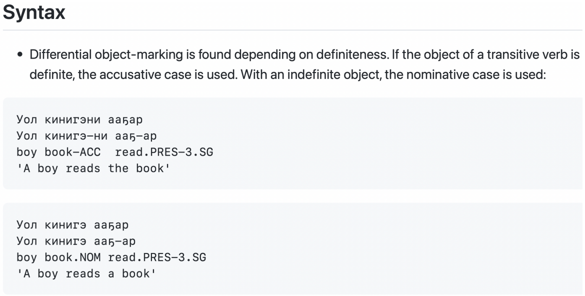
Besides Yakut, five other Turkic languages are represented in UD: Kazakh, Old Turkish, Tatar, Turkish (with nine treebanks), and Uyghur. A Kyrgyz treebank has been announced but has not yet been released. A comparison of Turkic treebanks in UD is given in Table 1. The presence of Old Turkish is important because it can shed light on diachronic processes within the Turkic family. Yet the disparity in the amount of sentences and tokens from one language to another is significant and calls for additional work before large scale analyses can be run on the set of several or all of the Turkic languages.

The annotation of the treebank is carried out based on the UD standards (Nivre et al., 2020), which use the CoNLL-U format[3]. The CoNLL-U file format requires the presence of ten columns: index, form, lemma, universal part-of-speech, language specific part-of-speech, morphological features, head, dependency relation, enhanced dependency graph, and allows for an optional additional annotation column. Although some columns only accept values from a pre-defined tagset, other columns can contain language specific features and values. For the Yakut treebank we carefully considered the terminology based not only on descriptions of Yakut, but also on more recent typological works and descriptions of other Turkic languages, especially the comparative ones (Deny et al., 1959; Johanson, 2021; Vinokurova, 2005). This decision allows researchers to grasp similar features of the Turkic languages more readily when working with the treebank.

Language	Sentences	Tokens
Kazakh	1.078	10.383
Old Turkish	18	221
Tatar	66	1.119
Turkish	> 50.000	> 500.000
Uyghur	3.456	40.236
Yakut	96	495

Table 1: Turkic languages in UD and the current state of their treebanks. The counts for Turkish are from all nine treebanks taken together.

The standardized documentation for features and their respective values as well as for dependency relations which are able to account for language specific constructions is a not only a useful reference but an essential step in developing NLP resources. An example of documented features in the current version of the Yakut treebank[4] is given in Figure 2 below. The full documentation can be accessed on the treebank hub page[5].



Syntax

- Differential object-marking is found depending on definiteness. If the object of a transitive verb is definite, the accusative case is used. With an indefinite object, the nominative case is used:

```
Уол кинигэни ааҕар
Уол кинигэ-ни ааҕ-ар
boy book-ACC  read.PRES-3.SG
'A boy reads the book'
```

```
Уол кинигэ ааҕар
Уол кинигэ ааҕ-ар
boy book.NOM read.PRES-3.SG
'A boy reads a book'
```

Figure 2: Documentation of a syntactic feature from the Yakut UD-treebank.

The competitive scores reached in the ConLL 2017 and 2018 Shared Tasks, illustrate the suitability of the UD framework for the development of high-accuracy parsers and other downstream NLP tasks (Zeman et al., 2018). It is based on the documentation of the features that the morphological analyzer is being built.

[2] https://universaldependencies.org/guidelines.html.

[3] https://universaldependencies.org/format.html.

[4] https://github.com/UniversalDependencies/docs/blob/pages-source/_sah/index.md.

[5] https://universaldependencies.org/treebanks/sah_yktdt/index.htm

2.1. The Annotation Process

Since Yakut has, since 1939, an official orthography, all texts available are written in it, which consists of the Cyrillic alphabet with five additional letters (Menz and Monastyrev, 2022). Some of the letters in the Russian alphabet are used exclusively in foreign words. As a consequence of this orthography, texts do not require pre-processing of transcription.

At present, only manual annotation is being carried out by these authors (TM and FFG)[6]. Supervised computational methods for the annotation are not yet possible due to the low amount of annotated sentences to be used as a training set. Once a few hundred sentences will have been manually annotated it will be possible to employ the UD-Pipe (Straka, 2018) to speed up the annotation process. This tool represents a trainable pipeline for processing CoNLL-U format, POS tagging, lemmatization, tokenization, and parcing. With ever growing training set, the growth of the treebank will thus also accelerate since expert judgment will be needed mostly for checking and correcting any erroneous tags made by the algorithm. So far transfer approaches have not been considered due to the small amount of annotated sentences.

3. Morphological analyzer

Morphological analysis is a basic component for a large number of automatic text processing systems, including machine translation, POS tagging, information retrieval, and information extraction. The effectiveness of the morphological analyzer largely depends on the effectiveness of all its subsequent stages.

The Yakut analyzer is being built based on data from (Kirov et al., 2018) with POS being extended manually. We are using a finite-state compiler Foma (Hulden, 2009), which is based on lexicon and rules. The lexicon stores a list of words to which morphological analysis is applied. The rule transducers are established from regular expressions and applied to the list of identified word forms. The rules are manually defined based on specialized literature and on native speakers judgement. Currently, approximately twenty rules have been implemented only regarding nouns and verbs. We suspect that with a couple hundred rules some meaningful results could be obtained.

For the system to perform better we need to have a large lexical database since the greater the number of unique word forms, the higher the accuracy of the morphological analysis. Therefore, we use the wordset for Yakut provided by the Universal Morphology project (UniMorph) (Kirov et al., 2018). UniMorph offers lists with lemmas and universal feature schemas with morphological categories. In the Yakut data nearly 600.000 different word forms were identified pertaining to almost 6.000 lemmata.

The morphological analyzer we are building for Yakut interacts with the morphological features and values on the Yakut treebanks, as exemplified in Figures 3, 4, and 5.

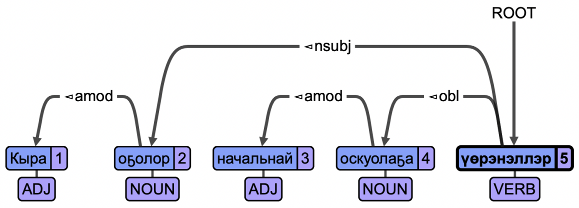
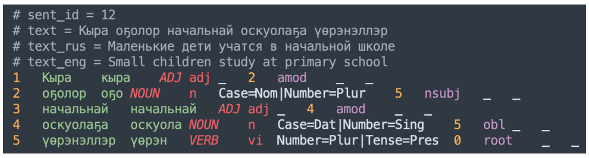
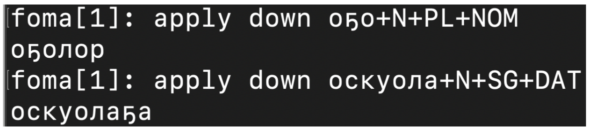


Figure 3: Example of dependency annotation from the Yakut UD-treebank.



```
# sent_id = 12
# text = Кыра оҕолор начальнай оскуолаҕа үөрэнэллэр
# text_rus = Маленькие дети учатся в начальной школе
# text_eng = Small children study at primary school
1   Кыра    кыра    ADJ adj _   2   amod    _   _
2   оҕолор  оҕо NOUN    n   Case=Nom|Number=Plur    5   nsubj   _   _
3   начальнай   начальнай   ADJ adj _   4   amod    _   _
4   оскуолаҕа   оскуола NOUN    n   Case=Dat|Number=Sing    5   obl _   _
5   үөрэнэллэр  үөрэн   VERB    vi  Number=Plur|Tense=Pres  0   root    _   _
```

Figure 4: Example of dependency annotation in CoNLL-U format from the Yakut UD-treebank.



```
foma[1]: apply down оҕо+N+PL+NOM
оҕолор
foma[1]: apply down оскуола+N+SG+DAT
оскуолаҕа
```

Figure 5: Example of network generation using Finite-State transducer.

Unfortunately, at this point, initial stage, we cannot evaluate the analyzer. A test-set is being prepared along the increment of rules.

4. Conclusion

We have briefly introduced the Yakut UD-treebank and the Yakut morphological analyzer that we intend to complete by the end of the year. Although we are still at an initial phase of the project, its presentation intends to spread information on the Yakut language an motivate the development of other treebanks, morphological analyzers, and lend support to the UD framwork so that more under-represented languages might profit from it and build on the existing set of data and tools.

Future work will focus on improving the precision and coverage of the morphological analyzer. A sequence-to-sequence recurrent neural network model (Sutskever et al., 2014) which produces morphological analysis for given text as output is also

[6] Both authors are computational linguists. Tatiana Merzhevich has some command of Sakha.

planned. Future work should also seek a closer interaction with tools for other Turkic languages, which as a consequence could enable profit from Yakut tools. While aware that there is a long path ahead, we look forward to receiving suggestions and engaging with the NLP community through this work since we believe that such interaction is essential and results in more robust and user-friendly resources.

5. Acknowledgements

The research presented in this paper is supported by the by European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 834050).

6. Bibliographical References

Anonymous. (2012). Sakhatyla (online sakha-russian / sakha-english dictionary. https://sakhatyla.ru.

De Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal dependencies. *Computational linguistics*, 47(2):255–308.

Deny, J., Grønbech, K., Scheel, H., and Velidi, T. Z. (1959). *Philologiae turcicae fundamenta*. Aquis Mattiacis apud Franciscum Steiner.

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2021). *Ethnologue: Languages of the World. Twenty-fourth edition*, volume 16. SIL international, Dallas, TX.

ELP. (2020). Endangered Languages Project: Catalogue of endangered languages. http://www.endangeredlanguages.com.

Gerdes, K., Guillaume, B., Kahane, S., and Perrier, G. (2018). SUD or surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium, November. Association for Computational Linguistics.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *EACL*.

Johanson, L. (2021). *Turkic*. Cambridge Language Surveys. Cambridge University Press.

Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Mielke, S. J., McCarthy, A., Kübler, S., Yarowsky, D., Eisner, J., and Hulden, M. (2018). UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Menz, A. and Monastyrev, V. (2022). Yakut. In Lars Johanson et al., editors, *The Turkic languages*, chapter 29, pages 444–460. Routledge, 2 edition.

Moseley, C. (2010). *Atlas of the World's Languages in Danger*. UNESCO, 3 edition.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection.

Nordhoff, S. and Hammarström, H. (2011). Glottolog/langdoc: Defining dialects, languages, and language families as collections of resources. In *Proceedings of ISWC 2011*.

Osborne, T. and Gerdes, K. (2019). The status of function words in dependency grammar: A critique of universal dependencies (ud). *Glossa: a journal of general linguistics*, 4(1):1–28.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium, October. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

Vinokurova, N. (2005). *Lexical categories and argument structure: A study with reference to Sakha*. Ph.D. thesis, Utrecht University. Unpublished PhD thesis.

WiNLP 2021 Workshop. (2021). *A Prototype Free/Open-Source Morphological Analyser and Generator for Sakha*. EMNLP. https://github.com/apertium/apertiumsah.

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018). CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium, October. Association for Computational Linguistics.