YakuTools
Yakut Treebank and Morphological Analyzer

Tatiana Merzhevich and Fabrício Ferraz Gerardi

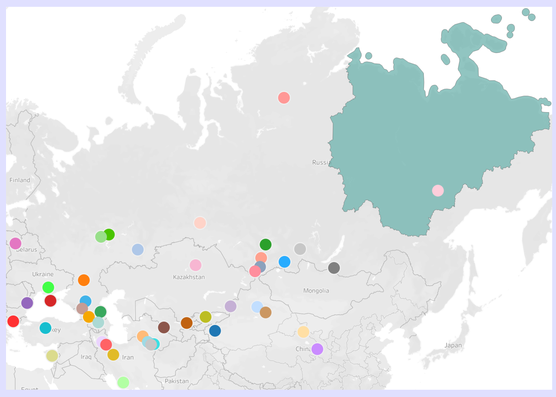
Universität Tübingen, Seminar für Sprachwissenschaft
{tatiana.merzhevich,fabricio.gerardi}.uni-tuebingen.de

Introduction

We present 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.

The publication of both the treebank and the analyzer serve this purpose with the prospect of evolving into a benchmark for the development of NLP tools (for other Turkic languages as well).

Sakha Republic



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.

The Yakut Language

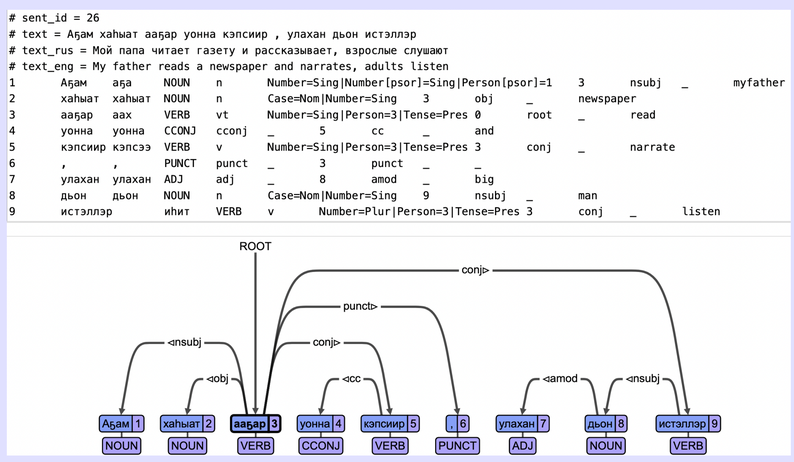
Yakut or Sakha (ISO sah, Glottocode yaku1245) is the easternmost member of the Turkic language family, spoken in the Republic of Sakha (Yakutia). In 2021 estimated population of the republic was about 1 million people. Of these, the half is ethnic Yakuts.

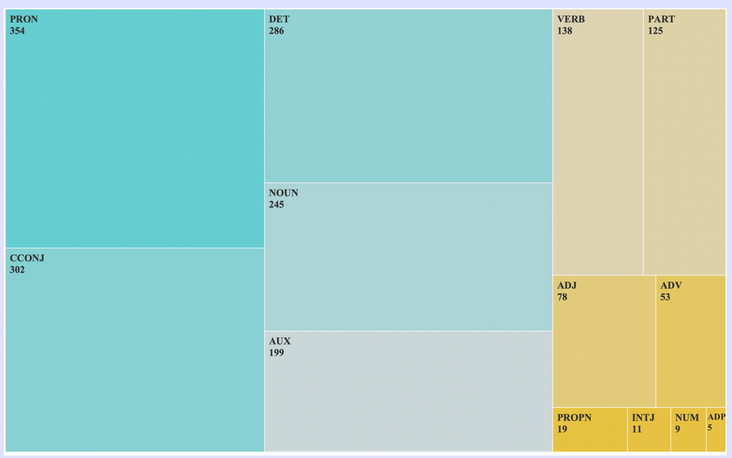
Inspite the number of speakers, the language is considered as threatened by Glottolog and Ethnolog. Yakut is an under-represented language whose prominence can be raised by making reliably annotated data and NLP tools that could process it freely accessible.

Yakut UD-Treebank

The annotation of the treebank is carried out based on the UD standards, which use the CoNLL-U format. 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. Currently, the corpus contains 96 sentences, 495 tokens and 496 syntactic words.

```
# sent_id = 26
# text = Аҕам хаһыат ааҕар уонна кэпсиир , улахан дьон истэллэр
# text_rus = Мой папа читает газету и рассказывает, взрослые слушают
# text_eng = My father reads a newspaper and narrates, adults listen
1       Аҕам    аҕа     NOUN    n       Number=Sing|Number[psor]=Sing|Person[psor]=1    3       nsubj   _       myfather
2       хаһыат  хаһыат  NOUN    n       Case=Nom|Number=Sing    3       obj     _       newspaper
3       ааҕар   аах     VERB    vt      Number=Sing|Person=3|Tense=Pres 0       root    _       read
4       уонна   уонна   CCONJ   cconj   _       5       cc      _       and
5       кэпсиир кэпсээ  VERB    v       Number=Sing|Person=3|Tense=Pres 3       conj    _       narrate
6       ,       ,       PUNCT   punct   _       3       punct   _       _
7       улахан  улахан  ADJ     adj     _       8       amod    _       big
8       дьон    дьон    NOUN    n       Case=Nom|Number=Sing    9       nsubj   _       man
9       истэллэр        иһит    VERB    v       Number=Plur|Person=3|Tense=Pres 3       conj    _       listen
```





Example of annotation (CoNLL-U and visualization (left). Treemap for the amount of POS from the current version of the treebank.

Morphological Analyzers

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.

Yakut morphological analyzer is being built based on the following approaches with trained data provided from the Yakut UD-treebank:

- a neural baseline encoder-decoder model
- a manual rules extraction using finite-state tools (FST)

Data-Driven Approach

The first Neural Network based morphological analyzer for Yakut is implemented on a basis of a character-level sequence-to-sequence LSTM model which decodes word forms into lemmas with POS-tags. Seq2Seq models have been applied successfully in the field of Morphological Reinflection where they secured first places in the SIGMORPHON shared tasks.

```
Input word(FORM): уочаратынан
Output(LEMMA+POS): уочарат<NOUN><Number=Sing><Case=Ins>

- - -
Input word(FORM): куоракка
Output(LEMMA+POS): куорат<NOUN><Number=Sing><Case=Dat>

- - -
Input word(FORM): буруйум
Output(LEMMA+POS): буруй<NOUN><Number=Sing><Case=Nom><Number[psor]=Sing><Person[psor]=1>

- - -
Input word(FORM): нууччалар
Output(LEMMA+POS): нуучча<NOUN><Number=Plur><Case=Nom>
```

Example of an encoder-decoder output.

The model accuracy for nouns achieved **94%**.

Rule Based Approach

Finite State Transducers (FST) are a popular method for morphological analysis. These tools have proven to achieve high lexical coverage and accuracy. We are using a finite-state compiler Foma, which is based on lexicon and rules.

Lang	all acc both	lemma	feats	unseen	#total	#both	#lemma	#feats	#unseen
sah	62.745 62.745	0.000	0.000	0.000	153	153	0	0	0

FST accuracy score for Yakut nouns.

The accuracy of FST methods using Foma for nouns in Yakut UD-treebank has reached **62%**.

Once a few hundred sentences will have been manually annotated it will be possible to employ the UD-Pipe to speed up the annotation process.

References

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

Hammarström, H., Forkel, R., Haspelmath, M., Bank, S. (2022). Glottolog 4.6. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.6578297 (Available online at http://glottolog.org, Accessed on 2022-06-15.)

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

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