A Prototype Free/Open-Source Morphological Analyser and Generator for Sakha

Sardana Ivanova
Helsingin yliopisto
Helsinki, 00014 Suomi
sardana.ivanova@helsinki.fi

Francis M. Tyers
Indiana University
Bloomington, IN 47405 USA
ftyers@iu.edu

Jonathan N. Washington
Swarthmore College
Swarthmore, PA 19081 USA
jonathan.washington@swarthmore.edu

1 Introduction

This paper describes the development of a morphological transducer for Sakha using free/open-source tools available as part of Helsinki Finite State Technology (HFST). Sakha is a Turkic language, with around 450 000 native speakers, primarily residing in the Sakha Republic. The Sakha Republic is located in Northeast Asia, and is part of the Russian Far East. Sakha speakers are subject to increasing economic (Streletskiy et al., 2019) and cultural (Lavrillier and Gabyshev, 2021) peril due to climate change.

The transducer described in this paper provides morphological analysis and generation for Sakha, is entirely hand-crafted, and is publicly available under the GPL v3 Free/Open Source licence.[1] Morphological transducers can be used in a wide range of language technology applications and "downstream tasks"; e.g., they may be repurposed as spell checkers and used in rule-based machine translation pipelines. The Sakha transducer described here is currently used in Revita, a language learning application designed to support individual efforts at language maintenance (Katinskaia et al., 2018; Ivanova et al., 2019). Morphological transducers are an important technology for NLP, since they are linguistically informed and well understood, and require a single development cycle (Butt, 2020).

2 Background and Methodology

Despite being rather divergent from other Turkic languages, Sakha shares a lot of properties with them: it can be described as agglutinating, meaning words may be inflected using a series of affixes; the word order is generally Subject-Object-Verb; and there are backness and rounding vowel harmony systems.

The function of a morphological transducer is twofold: morphological generation takes surface forms (e.g., `атын`) as input and returns all possible lexical forms (e.g., `ат<n><px3sg><acc>/атын<adj>`, cf. /ɑt/ 'horse', /ɑtɯn/ 'different'), and morphological analysis takes lexical forms (e.g., `ат<n><px3sg><acc>`) and returns one or more surface forms (e.g., `атын`). Morphological transducers are implemented as finite state transducers (FSTs), and in this case are compiled from hand-coded lexical, morphotactic, and morphophonological generalisations. The lexicon and inflectional morphotactics are encoded in the `lexc` formalism, and the morphophonology in `twol`; the two files are compiled as FSTs using Helsinki Finite-State Technology (HFST) (Lindén et al., 2011), and these FSTs are intersected to produce the full transducer, per Koskenniemi (1983) and Beesley and Karttunen (2003). This follows the methodology used in previous Turkic FSTs (Washington et al., 2019).

Stems were added to the lexicon by their frequency of forms containing them in the Sakha Wikipedia corpus. Specifically, we went through an iterative process (documented in Washington et al. (2016)) of analysing the corpus using the transducer, identifying the stems of the most frequent unrecognised forms, adding those, recompiling, and running analysis again. Some examples of non-lexical development follow.

3 Contents and design

The transducer currently includes over 10 500 stems, consisting of around 5 400 nouns, over 2 100 proper nouns, over 1 300 adjectives, and over 1 000 verbs. The remaining stems are divided between adverbs, numerals, pronouns, postpositions, conjunctions, and determiners. The tagset consists of 105 separate tags, 15 covering the

[1] `https://github.com/apertium/apertium-sah`

main parts of speech (noun, adjective, verb, adverb, postposition, etc.) and 90 covering lexical subcategory—e.g., transitivity, proper noun class, determiner type, etc.—and morphological function—e.g., case, number, person, possession, tense-aspect-mood, etc. The tags are based on the Apertium tagset[2] defined in the `lexc` source as multi-character symbols, between less than '<' and greater than '>' symbols, along with comments describing their usage.

Because long vowels in Sakha (`{I}{I}`, `{A}{A}`) behave as their short-vowel counterparts with regard to vowel harmony, but `{I}{A}` (high+low vowel) diphthongs behave like high vowels and not like low vowels (they round after any round vowel, and do not trigger the rounding of low vowels), each `twol` harmony rule (single character-to-character mappings) had to be sensitive to whether a harmonising vowel character is a component of a long vowel or a diphthong or not, and many of the alternations required multiple rules to implement.

Sakha also has consonant assimilation processes that apply in both directions. The verb form /tutn-bIt-A/ 'use–PAST–3' is realised as *тутуммута* [tutummuta], where the /n/ triggers nasalisation of the following /b/, and the /b/ triggers labialisation of the preceding /n/. In `twol`, this sort of mutual influence is not problematic, and may be implemented simply as two rules sensitive to the underlying form of adjacent consonants.

Stem alternations like *уһун* 'swim.IMP' / *үстар* 'swim-PRES' exhibit three single-character alternations in sequence, here: h 'h'/c 's' due to intervocalic lenition, harmonised high vowel/∅ due to consonant cluster restrictions, and н 'n'/т 't' due to sonority restrictions. Each of these alternations required at least one `twol` mapping, variably sensitive to the other alternations and to other parts of the morphophonological context. The morphological forms were implemented using a special archiphoneme `{y}` that was conditioned to alternate between nothing and a context-appropriate high vowel, as in previous Turkic transducers (Washington et al., 2019).

Sakha exhibits a high number of non-finite verb forms, many of which have finite uses as well. Previous grammars like Убрятова et al. (1982) categorise these forms as either participles or converbs, and do not present a detailed categorisation of their uses. As part of the construction of this transducer, we identified for each of these non-finite verb forms whether it had finite, verbal noun, verbal adjective, verbal adverb, or infinitive uses, and implemented each use separately. We found that while this results in some syncretism for many forms, there is not a strict participle/converb binary as presented in previous sources. This work, documented in more detail in Washington and Tyers (2019; Washington et al. (2022), constitutes a novel understanding of Sakha grammar.

4 Evaluation

We calculated the naïve coverage (percentage of forms receiving an analysis, whether correct or not) of the analyser on a recent version of the Sakha Wikipedia (~2.59M tokens) and a newspaper corpus (>16M tokens) (Leontiev, 2015). Coverage on the Wikipedia corpus is 88.76%. Most top unknown tokens are foreign words (Russian and English words not used in Sakha) and pieces of URLs. Coverage on the newspaper corpus is 87.68%. The slightly higher coverage on the Wikipedia corpus despite the foreign content is expected, since we used frequency lists from this corpus to add stems to the transducer. The mean ambiguity of the transducer (average number of analyses returned by the transducer per token) over the Wikipedia corpus is 2.32 and over the newspaper corpus is 2.22.

To measure the transducer's precision (percentage of analyses returned by the transducer that is correct) and recall (percentage of the correct analyses that is returned), we selected 500 valid words of Sakha randomly from the Wikipedia corpus, ran them through the analyser, and manually annotated this list to create a gold standard. Manual annotation consisted of adding analyses, and removing and correcting returned analyses for each form. The gold standard was then compared to the original list of analyses returned by the analyser. Precision was 99.9%, recall was 69.9%; i.e., nearly every form returned by the transducer was deemed correct, but many correct analyses were not returned by the transducer (mostly due to low coverage).

5 Future work

The main future work to be done to improve this transducer is increasing the size of the lexicon. While a good level of coverage has been achieved with only around 10 000 stems, production-quality morphological analysers have tens of thousands of stems, even for morphologically-rich languages like Sakha. A number of minor issues in the implementation of some morphophonological alternations in the transducer were identified recently when

[2] `https://wiki.apertium.org/wiki/Tagset`

the transducer was used as part of data generation for a shared task (Pimentel et al., 2021). It is expected that coverage will increase as a result of fixing these issues as well. Once good coverage has been achieved with a morphological analyser, the next logical step is to start work on morphological and syntactic disambiguation. As the mean ambiguity figures suggest, there is a lot of work that can be done on disambiguation.

6 Conclusion

We have presented, to our knowledge, the first ever published morphological analyser and generator for Sakha, a marginalised language of Siberia. The transducer has coverage of between 87-89%, and high precision. In the development of the analyser, we have expanded linguistic knowledge about Sakha, and developed strategies for complex grammatical patterns. The transducer is already being used in downstream tasks, including computer assisted language learning applications for linguistic maintenance and computational linguistic shared tasks.

References

Kenneth R. Beesley and Lauri Karttunen. 2003. Two-level rule compiler. `https://web.stanford.edu/~laurik/.book2software/twolc.pdf`.

Miriam Butt. 2020. Building resources: Language comparison and analysis. Invited talk at *The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sardana Ivanova, Anisia Katinskaia, and Roman Yangarber. 2019. Tools for supporting language learning for Sakha. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa'19)*. Linköping University Electronic Press.

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2018. Revita: a language-learning platform at the intersection of ITS and CALL. In *Proceedings of LREC: 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan.

Kimmo Koskenniemi. 1983. *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*. Number 11. University of Helsinki Department of General Linguistics, Helsinki.

Alexandra Lavrillier and Semen Gabyshev. 2021. An indigenous science of the climate change impacts on landscape topography in siberia. *Ambio*.

Nyurgun Leontiev. 2015. The newspaper corpus of the Yakut language. In *Proceedings of TurkLang 2015*, page 233.

Krister Lindén, Erik Axelson, Sam Hardwick, Tommi A Pirinen, and Miikka Silfverberg. 2011. Hfst—framework for compiling and applying morphologies. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 67–85. Springer.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. Sigmorphon 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259. Association for Computational Linguistics.

Dmitry A Streletskiy, Luis J Suter, Nikolay I Shiklomanov, Boris N Porfiriev, and Dmitry O Eliseev. 2019. Assessment of climate change impacts on buildings, structures and infrastructure in the russian regions on permafrost. *Environmental Research Letters*, 14(2).

J. N. Washington and F. M. Tyers. 2019. Delineating Turkic non-finite verb forms by syntactic function. In *Proceedings of the Workshop on Turkic and Languages in Contact with Turkic 4*, pages 132–146.

J. N. Washington, A. Bayyr-ool, A. Salchak, and F. M. Tyers. 2016. Development of a finite-state model for morphological processing of tuvan. *Родной Язык*, 1(4):156–187.

Jonathan Washington, Ilnar Salimzianov, Francis M. Tyers, Memduh Gökırmak, Sardana Ivanova, and Oğuzhan Kuyrukçu. 2019. Free/open-source technologies for Turkic languages developed in the Apertium project. In *Proceedings of TurkLang 2019*.

Jonathan N. Washington, Francis M. Tyers, and Ilnar Salimzianov. 2022. Non-finite verb forms in Turkic exhibit syncretism, not multifunctionality. *Special volume on multifunctionality and syncretism in non-finite forms*.

Е. И. Убрятова, Е. И. Коркина, Л. Н. Харитонов, and Н. Е. Петров, editors. 1982. *Грамматика современного якутского литературного языка: Фонетика и морфология [E. I. Ubryatova et al. Grammar of the modern Yakut literary language: Phonetics and morphology]*. Москва: Наука.