Challenges in development of corpora of Samoyedic languages from legacy data

Maria Brykina, Josefina Budzisch, Aleksandr Riaposov, Alexandre Arkhipov
University of Hamburg

In this paper, we're dealing with corpora of endangered Samoyedic languages, based primarily on legacy data of various kinds: INEL Kamas, Selkup and Nenets corpora and Selkup Language Corpus (SLC). We will outline a few problems that arose in the process of their creation. Most challenges are due to the heterogeneity of the data as well as to the many open questions in the grammar and lexicon of the languages in focus. Source materials within each corpus vary in dialects and/or the time period represented, in modality (written vs. audio), in transcription conventions used in different sources, or in combination of these.

Kamas data have two main sources: Kai Donner's written collection (1912–1914) and the audio recordings of the last speaker, Klavdiya Plotnikova, by Ago Künnap and others in 1963–1970. Selkup data in two corpora come from several publications and from the fieldwork archive of Angelina Kuzmina who worked in many places where the Selkup lived in 1960s–1970s. Nenets data come from several publications and from the fieldwork archive of Svetlana Burkova.

Challenges in transcription. Diverse orthographies and transcription conventions used in different sources for each language or major variety need first to be converted into a common representation in order to make uniform morphological analysis possible. This was chosen to be a Latin-based transcription, similar though not identical across different corpora. Conversion rules are created based on prior knowledge of phonology of each variety and applied automatically, with further manual adjustments when needed. Sometimes, however, our knowledge evolves during the text analysis, thereby causing the need to make changes to already analyzed data.

Challenges raised by alternative text versions. Although the Kamas texts collected by Kai Donner are best known from their posthumous edition by A. Joki (1944), their phonologized transcription used as basis for glossing relies on G. Klumpp's unpublished edition of Donner's manuscripts. The latest version of the corpus also includes the narrow transcription by Donner from the same unpublished edition. Meanwhile, the original German translation by A. Joki as published in (Joki 1944) is also provided, supplemented by an edited modern translation based on the new reading.

In the Selkup archive of Angelina Kuzmina, the same text recorded from the same speaker is sometimes found both as an audio recording and written down by the collector. Only in two cases the audio and the written version were judged close enough to be treated as the same text

item. The remaining more than two dozen pairs had to be analyzed and included in the corpus as different text versions sharing a similar storyline and overlapping to some extent.

Challenges in lemmatization. The absence of a formalized standard, combined with diverse transcribing traditions and a high dialectal diversity (in case of Selkup and Nenets), results in significant variation across the data. This makes it difficult to identify a single "canonical" form for a word or morpheme. Consequently, choosing consistent lexical and grammatical representations remains a major challenge in the development of these corpora. E.g., forms encountered across Selkup dialects for the noun 'grandmother' include *imɨl'a, imn'a, nima, ni:ba, al'diga, al'ǯiga, ad'uka*; for 'horse' — *čünd(V), č'un(n)V, č'untV, kVn(n)V, kVtdV, kVnd(V),* among others.

Challenges in morphological analysis. In the case of lesser studied languages, a thorough analysis of texts typically encounters unknown morphological elements. Particularly in Samoyedic languages with their rich verbal morphology, derivational suffixes that have not yet been described or classified frequently emerge. To address this, placeholder labels such as DRV ("unspecified derivation") are used in the corpora. This approach allows marking unidentified morphemes systematically, with the goal of future classification and analysis, which can be hindered by small numbers of examples in the corpus. E.g. out of a few dozens of such unidentified derivation markers in a working version of the Forest Nenets corpus, only eight had 10 or more occurrences.

References

Brykina, Maria; Orlova, Svetlana; Wagner-Nagy, Beáta. 2021. *INEL Selkup Corpus*. Version 2.0. Publication date 2021-12-31. Archived at Universität Hamburg. https://hdl.handle.net/11022/0000-0007-F4D9-1.

Budzisch, Josefina; Harder, Anja; Wagner-Nagy, Beáta. 2019. *Selkup Language Corpus (SLC).* Archived in Hamburger Zentrum für Sprachkorpora. Version 1.0.0. Publication date 2019-02-08. http://hdl.handle.net/11022/0000-0007-D009-4.

Budzisch, Josefina; Wagner-Nagy, Beáta. 2024. INEL Nenets Corpus. Version 1.0. Publication date 2024-12-31. https://hdl.handle.net/11022/0000-0007-FE37-E. Archived at Universität Hamburg.

Gusev, Valentin; Klooster, Tiina; Wagner-Nagy, Beáta. 2023. *INEL Kamas Corpus.* Version 2.0. Publication date 2023-12-31. http://hdl.handle.net/11022/0000-0007-FC25-4. Archived at Universität Hamburg.

Joki, Aulis Johannes. 1944. *Kai Donners Kamassisches Wörterbuch nebst Sprachproben und Hauptzügen der Grammatik*. Helsinki: Suomalais-Ugrilainen Seura.