Linguistic variation and minor languages corpora: a case study of Mansi dialects

Daria Zhornik, Fyodor Sizov

Moscow State University; Institute of Linguistics, Russian Academy of Sciences

Introduction

- Minor languages corpora exhibit a high level of variation, as these languages lack a written norm (and often any written practice at all).
- Gathering such data and integrating this material into a readily usable corpus are equally non-trivial tasks with many pitfalls and challenges.
- We face a very limited number of generally accepted authoritative computer tools for this type of linguistic data, though some efforts in this direction are worth noticing (cf., for example, [Simon & Mus 2017]).
- No tolerable resource with annotated texts from Mansi dialects either.
- Our project on Mansi documentation: www.digital-mansi.com.
- One of our goals: to create a multidialectal multimedia corpus of all Mansi dialects, for which at least some documentation exist; NB: variation seems the main problem.
- The corpus contains highly diverse data (dialectal variation, heterogeneity of writing systems, different time periods etc.)

The Mansi language

- Mansi < Ob-Ugric < Finno-Ugric < Uralic
- An indigenous language of Western Siberia, 2010 census: 938 speakers
- 4 dialectal groups (Northern, Western, Eastern, Southern), but today only some Northern dialects survive.
- Documentation of the Upper Lozva dialect: field trips in 2017-2018, see [Zhornik, Pokrovskaya 2017]
- Upper Lozva Mansi: Northern part of the Sverdlovsk oblast with approx. 100 speakers.

Previous research

- Compared to other Uralic languages, digitized Mansi texts appear rather scarcely, see [Horváth et al. 2017] for description.
- The websites of the newspaper "Lūima Sēripos" (http://www.khanty-yasang.ru/luima-seripos) and of the Ob-Ugric institute of applied research and development (https://ouipiir.ru); NB: only plain texts (no glosses!)
- A corpus of 272 annotated (to some extent) Mansi texts as part of the Ob-Ugric Database (OUDB): Elena Skribnik, Munich, http://www.babel.gwi.uni-muenchen.de, [Schön, Wisiorek 2016].
- "Languages under the influence" project [Simon, Mus 2017], however, this corpus is currently not available and not open for evaluation.
- The existing tools are poorly adapted for tackling variation of various types, as they rarely include dialectal texts.
- This objective is, however, central to our project aimed at the extensive corpus documentation of the Mansi linguistic continuum.

Corpus

- The corpus presented on the website (www.digital-mansi.com/corpus) is going to consist of a significant number of subcorpora (currently under construction) distributed according to dialectal areas.
- Several subcorpora will also feature standard written Mansi: newspapers, fiction, Bible translations, and so forth.
- Each subcorpus (or collection thereof) exhibits its own lexicon, morpheme list and grammar rules.
- The written texts for the corpus originate from different sources:
  - books and newspapers in standard Mansi
  - fieldwork data from various Mansi dialects recorded in the 19th and 20th centuries by Arturi Kannisto, Antal Reguly, Béla Munkácsi, Nikolay Chernetsov etc.
  - Audio materials from the Upper Lozva dialect
- Only one multimedia subcorpus (provided with audio files), as most of the Mansi dialects are extinct and have no audio recordings whatsoever.
- The Upper Lozva recordings are being transcribed and segmented into clauses in ELAN and afterwards this markup is synchronized with morphological annotation.

Preliminary text processing

- For printed sources, we apply Optical Character Recognition (OCR) techniques; the Mansi module was specially developed and trained by the authors
- All Mansi texts use different writing systems depending on the author or editor
- The OCR software should be able to recognize numerous character sets and support various fonts.
- The texts which are supposed to be added to the corpus belong to 200 years long time span.
- During OCR, we have to decide which Unicode symbols correspond best to those found in texts
- OCR of standard Mansi texts: 84% accuracy.
- OCR processing for dialectal texts is still in progress, so the accuracy is yet unknown
- After OCR: a group of systematic mistakes produced by the engine is improved via regular replacements with the help of *sed* stream editor.
- Residuary mistakes may be found and corrected manually.

Morphological parser

- Available morphological parsers are mostly unable to cope with the task of variation processing.
- We have developed a special morphological analyzer **AmpEngine** based on the models borrowed from the universal morphological analyzer UniParser (see [Arkhangelskiy et al. 2012] for more detail).
- The main goal of AmpEngine is to capture cross-dialectal variation, as well as to integrate written and oral texts within the same automated system.
- AmpEngine is dictionary-based: drawing on all available Mansi dictionaries, we have created several databases (the current number of entries is 10262).
- The testing of the morphological parser was based on Mansi 7 newspaper articles (2317 wordforms)
- The results of parsing have been compared with the previously performed manual annotation, the latest version of the analyzer reached 87% accuracy.
- Typical mistakes: toponymes, calques or neologisms.
- The parsing example:

Тāн аквхурип тāрвитыӈ вāрмаль ōньщēгыт.

(Translation: They had the same difficult task.)

Тāн	аквхурип	тāрвитыӈ	вāрмаль	ōньщ-ēг-ыт
they(pl.)	the.same	difficult	task	have-PRS-3PL

sinew

Conclusions

- We believe that multidialectal corpora can serve as significant contribution to the field of corpus linguistics (especially when innovative methods are used to solve the problem of cross-dialectal variation), as well as to the documentation of under-described languages.
- The texts presented in our corpus cover the span of almost 200 years, so that the corpus allows us to perform diachronic studies and research the mechanisms of Mansi dialectal divergence.
- Our corpus is freely accessible on the Internet (http://digital-mansi.com/corpus), which allows scholars interested in either linguistic typology or Uralic languages gain access to a valuable collection of diverse Mansi texts.

References

1. Arkhangelskiy T., Belyaev O., Vydrin A. (2012), The creation of large-scale annotated corpora of minority languages using UniParser and the EANC platform, Proceedings of COLING 2012: Posters. Mumbai: The COLING 2012 Organizing Committee, ch. 9. p. 83–91.
2. Horváth C., Szilágyi N., Vincze V., Nagy A. Language technology resources and tools for Mansi: an overview. Proceedings of the Third International Workshop on Computational Linguistics for Uralic Languages, Saint-Petersburg.
3. Rombandeeva E. I. (2005), A Russian-Mansi dictionary [Russko-mansijskij slovar'], Mirall, Saint-Petersburg.
4. Simon E., Mus N. (2017), Languages under the influence: Building a database of Uralic languages, The 3rd International Workshop for Computational Linguistics of Uralic Languages, St. Petersburg, Russia, 23–24 January 2017, pp. 10–24.
5. Schön Zs., Wisiorek A. (2016), Ob-Ugric Database: Corpus and Lexicon Databases of Khanty and Mansi Dialects, Second International Workshop on Computational Linguistics for Uralic Languages in Szeged, Hungary.
6. Zhornik D. O., Pokrovskaya S. V. (2017), Documentation of the Upper Lozva Mansi dialect [Dokumentacija verhnelozvinskogo dialekta mansijskogo jazyka], Minor languages in big linguistics [Malye jazyki v bol'shoj lingvistike], Moscow State University, Moscow, Russia.