Toward a corpus of Tundra Nenets: stages and challenges in building a corpus

Nikolett Mus & Réka Metzger

Hungarian Research Institute for Linguistics

2–3 March 2021
4th Workshop on Computational Methods for Endangered Languages (online)

Background

- The Tundra Nenets corpus building work is carried out within the project titled "Theoretical and experimental approaches to dialectal variation and contact-induced change: a case study of Tundra Nenets" (FK_129235).

  ⇒ We seek traces of a contact-induced typological restructuring in two Tundra Nenets dialects having different sociolinguistic backgrounds.

  ⇒ We document, describe, and analyse the syntax and prosody of various types of questions.

Background (cont.)

- Tundra Nenets written texts were/will be collected to explore syntactic research questions, and to formulate working hypotheses.
- Tundra Nenets spoken data were/will be collected during fieldworks in the project period.
- Even though there are Tundra Nenets annotated witten and spoken texts available from the web, these sources primarily serve to sample the language.

$\Rightarrow$ Building a Tundra Nenets corpus by processing our archived materials would fill the gap in the currently available resources.

Outline

The Tundra Nenets language

The Tundra Nenets language

- Tundra Nenets is an endangered indigenous minority language.
- The language is spoken in three major administrative districts of the Russian Federation.

Figure 1: The Tundra Nenets-speaking area of the Russian Federation

The Tundra Nenets language (cont.)

- Tundra Nenets belongs to the Samoyedic branch of the Uralic family.
- The status of Tundra Nenets is 6b, i.e. *threatened*, on the EGIDS scale.
- There are c. 20,000 Tundra Nenets speakers.
- The language consists of three main dialectal groups: Western, Central, and Eastern.
- Tundra Nenets is exposed to different external and internal influences, and one can hardly find a Nenets speaker who is not bi- or multilingual.

The Tundra Nenets language (cont.)

- There is neither a unified literary language, nor a unified writing system of Tundra Nenets.
- There are online Tundra Nenets newspapers.
- A test version of Tundra Nenets Wikipedia is also present.
- Tundra Nenets videos and recordings are provided and archived by the Yamal Region broadcast.
- Tundra Nenets language tools, e.g. a text analyzer, paradigm and (number) word generators, and digital dictionaries, are available on the website of Giellatekno.

Methods and results

Workflow

- The workflow in Figure 2 shows the individual stages of our corpus-building work.

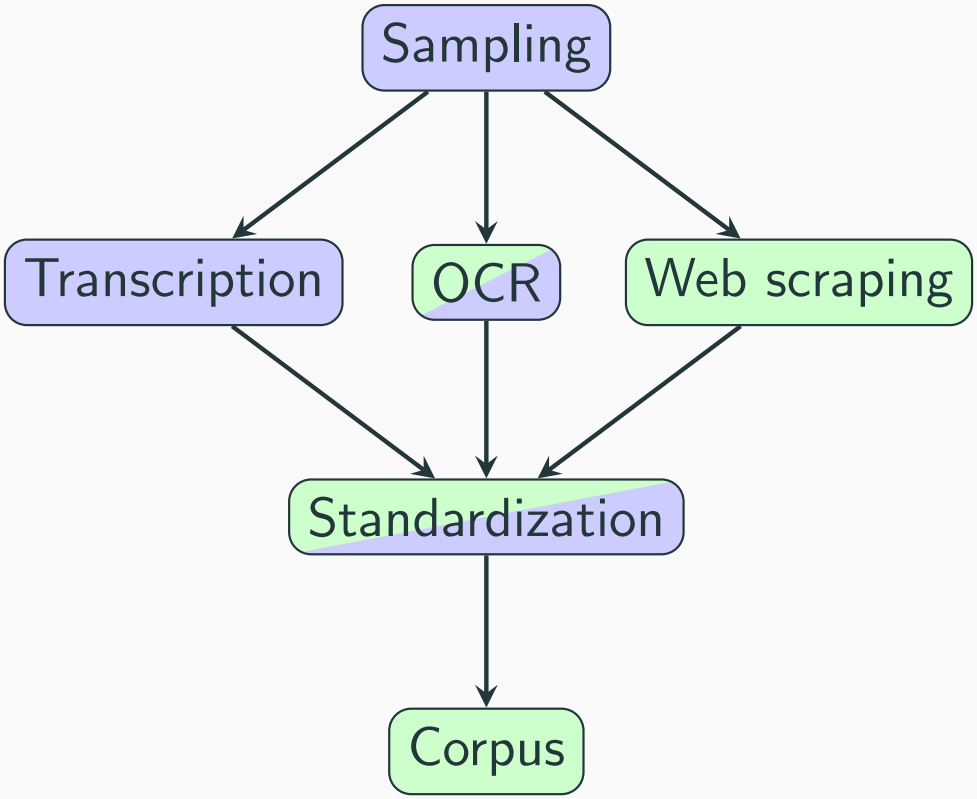

Figure 2: Workflow of corpus building process

Sampling

- We considered important the *reliability*, *naturalness*, *balancing* and *representativeness* of the various sampling ideals usually emphasized in the literature.

- Our intention was to equally represent:

  - the dialect(al group)s of the language,
  - the gender and age of the speakers,
  - and the spoken and written varieties of the language.

Sampling (cont.)

- Spoken texts were recorded by us and transcribed by our native speaker consultant (narratives).
- We classified our written sources on the basis of the reality of the speech event when the texts were created:
  - written texts that are potentially based on real situations (folklore, phrasebook, methodological handbook);
  - written texts produced in imagined situations (newspapers).

Text processing

- There were three types of texts we had to process in this stage: (i) recorded spoken texts, (ii) (published) written sources in print, (iii) written texts available from the web.

- We processed our data in different ways depending on their characteristics:

  - texts in (i) $\Rightarrow$ transcription,
  - texts in (ii) $\Rightarrow$ (scanning and) OCR,
  - texts in (iii) $\Rightarrow$ web scraping.

Text processing: transcription

- We recorded 5 Tundra Nenets texts in 2017.
- These texts were transcribed by our Tundra Nenets informant.
- The transcription of these texts is a simple orthographic transcription.

Text processing: OCR

- A significant amount of Tundra Nenets texts were available in print.
- We scanned these resources, and converted the .pdf files into .txt ones using the OCR software developed by Abbyy FineReader.
- We checked the accuracy of OCR two times.
  - We compared the input and output files manually.
  - We relied on some phonotactic constraints of Tundra Nenets. We created a wordlist and searched for forms that violates these phonotactic rules.

$\Rightarrow$ The output of this process was 243 Tundra Nenets texts.

Text processing: web scraping

- An online newspaper is available titled Няръяна вындер ('Red Tundra').
- We included the Tundra Nenets articles between 14/02/2013–18/04/2019.
- We implemented a web scraper in Python to make the harvesting process less time consuming.
- It collects URLs of the articles, then iterates through them and extracts the necessary metadata and data from the HTML tags using regular expressions.

  $\Rightarrow$ The output was 793 Tundra Nenets plain texts.

Standardization

- Standardization included cleaning the texts from extra white spaces and unifying the punctuation.
- Two encoding problems specific to Tundra Nenets occurred during this process:
    1. the same character was used with different (grammatical) functions,
    2. different characters/graphemes stood for the same phoneme.

Standardization (cont.)

1. The same character was used with different (grammatical) functions.

- The standard double quotation mark (U+0022) was used in quotations, and it also stands for a glottal stop phoneme in the texts.

  $\Rightarrow$ We inserted the quotes between French quotation marks (U+00AB, U+00BB).

Standardization (cont.)

2. Different characters/graphemes stood for the same phoneme.

- Both apostrophe (U+0027) and standard double quotation mark (U+0022) were used for the glottal stop phoneme

  $\Rightarrow$ We kept the original differentiation in the texts due to linguistics reason.

- The representation of the velar nasal phoneme in the texts: three different graphemes were used (U+04C9, U+04A2, U+04C8).

  $\Rightarrow$ After consulting with the community we use the one (U+04C8) that is standardly used in the virtual keyboard apps, e.g. in Gboard.

Corpus

- Our selection criteria of the corpus management system:
  - Powerful searching features even for non-annotated texts (simple and complex queries with regular expression).
  - The possibility of creating a parallel corpus.
  - The ability to create a corpus with an unsupported language.
  - Having a clear, user-friendly and customizable interface.

$\Rightarrow$ We use the open-source NoSketch Engine (NoSkE) corpus management system.

Corpus (cont.)

- NoSkE requires two files to be able to compile a corpus.
  - Vertical file: an XML file that contains all the texts vertically (every token and its metadata in a separate line).
  - Corpus configuration file: additional information takes place here such as language, encoding, description etc.
- We defined attributes in the XML file to make it possible to search by categories: id, source, genre, gender, dialect.

  $\Rightarrow$ NoSkE complied our Tundra Nenets (Monolingual) corpus with its 452,930 tokens.

Corpus (cont.)

- We created a Cyrillic keyboard to make the search easier.

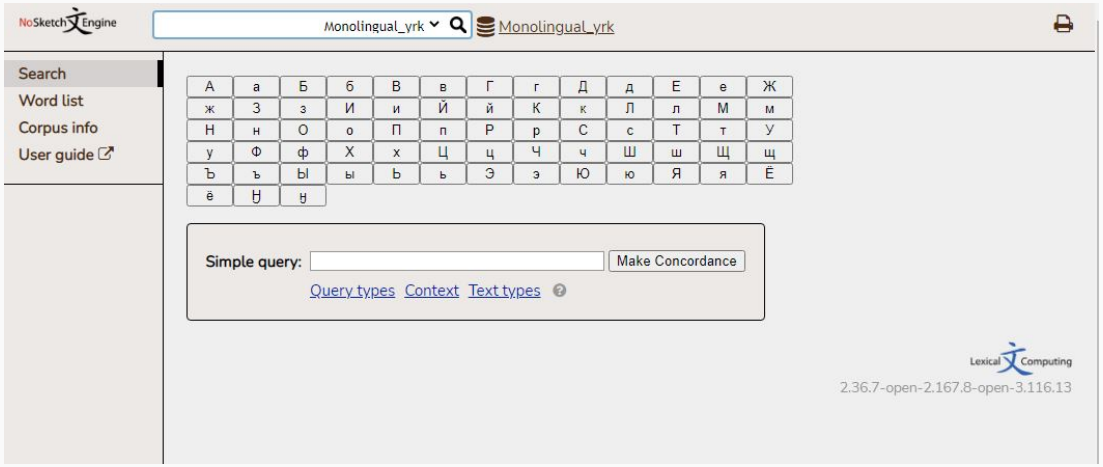

Figure 3: Tundra Nenets Monolingual Corpus

- We gathered detailed metadata of the texts: source, date of origin, type (written/spoken), (sub)genre, number of tokens and the name, gender, age and dialect of the speaker.

  $\Rightarrow$ This information is available in a catalogue from our website.

Ongoing work and future plans

Ongoing work and future plans

- We collect, process and include further texts.

  $\Rightarrow$ We contact archives and individual researchers, and participate in fieldworks.

- We create a parallel Tundra Nenets–Russian corpus: the Russian translations of the texts will be aligned at sentence level.

  $\Rightarrow$ Texts will be translated by Tundra Nenets–Russian bilingual speakers.

Ongoing work and future plans (cont.)

- We annotate the Tundra Nenets data at certain levels.
- We automatize some of the stages of our process described above.
- We contact researcher and speaker community in order to customize both the corpus and the user interface to their needs.

Thank you!

Acknowledgments

The support of the research project "Theoretical and experimental approaches to dialectal variation and contact-induced change: a case study of Tundra Nenets" (FK_129235) is gratefully acknowledged.