Archiving Tundra Nenets materials: towards designing a balanced Tundra Nenets corpus

Nikolett Mus (musn@nytud.hu) & Réka Metzger (metzger.reka@gmail.com)
Hungarian Research Institute for Linguistics

Introduction

This paper reports stages of an ongoing Tundra Nenets archiving and corpus building work. The main aims of our project are:

- to **collect** and **archive** written and spoken Tundra Nenets data
- to **make** these data **available** for the speaker and the researcher community
- to **support** (preferably synchronic syntactic) research on Tundra Nenets

The Tundra Nenets language

- 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 is spoken in three major administrative districts of the Russian Federation:
  the Nenets Autonomous Okrug
  the Yamalo-Nenets Autonomous Okrug
  the Taymyrsky Dolgano-Nenetsky District
- A few more groups of speakers can sporadically be found in the Khanty-Mansi Autonomous District, in the Komi Republic, and in the Murmansk region.
- There are three main dialect(al group)s of Tundra Nenets: Western, Central, and Eastern.
- The Russian language and culture has a great influence on the Tundra Nenets speaking community.
- The Tundra Nenets speakers are bi- or multilinguals.
- There is neither a unified literary language, nor a unified writing system of Tundra Nenets.

Background

Tundra Nenets sources available from the web primarily serve to sample the language, but cannot be considered as large, robust, balanced, or representative corpora.

Sampling

Our intention is to equally represent **dialect(al group)s** of the language, the **gender** and **age** of the speakers, the **date** of the recording, and both **spoken** and **written** varieties of the language.

→ We store these data in a catalogue (`https://tundranenetsdata.nytud.hu/`).

Fig. 1: The Tundra Nenets-speaking area of the Russain Federation

Archiving Tundra Nenets materials: towards designing a balanced Tundra Nenets corpus

Nikolett Mus (musn@nytud.hu) & Réka Metzger (metzger.reka@gmail.com)
Hungarian Research Institute for Linguistics

Our methods

Figure 2 shows stages of our corpus-building work.

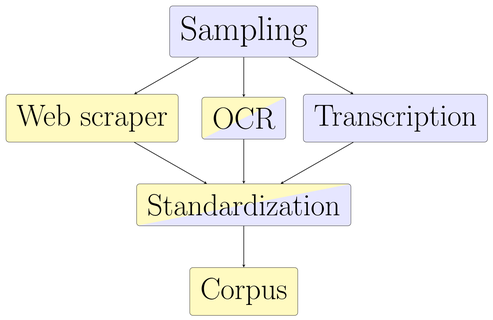

Fig. 2: Workflow of corpus building process

→ **Yellow** colour marks **automated** stages.

→ **Blue** colour indicates the works/tasks carried out **manually**.

Text processing

In order to convert the Tundra Nenets texts into machine-readable form, they were processed in different ways depending on their characteristics:

- written texts available from the web, e.g. newspapers ⇒ **web scraping**
- written sources in print, e.g. folklore compilations ⇒ **OCR**
- recorded spoken texts ⇒ orthographical **transcription**

The output of this stage is **UTF8** encoded **.txt** files.

Challenges during standardization

Encoding problems specific to Tundra Nenets to be solved:

- the **same character** used with **different** (grammatical) **functions**

  e.g. Standard double quotation mark used in quotations and for glottal stops. ⇒ The quotes were insterted between French quotation marks.

- **different characters**/graphemes stood for the **same phoneme**

  e.g. Apostrophe and standard double quotation mark used for glottal stops. ⇒ The original differentiation was kept due to linguistic reasons.

  e.g. Three representations of the velar nasal phoneme. ⇒ The one standardly used in the virtual keyboard apps (e.g. in Gboard) was kept.

Corpus

- The open-source NoSketch Engine (NoSkE) corpus management system is used.
- The corpus has not annotated or lemmatized yet.
- Tundra Nenets Monolingual corpus contains 452,930 tokens so far.
- Tundra Nenets Monolingual corpus is available on the website `https://tundranenetsdata.nytud.hu/bonito`

How to use the corpus?

- NoSkE offers simple and complex searches with **regular expressions**.
- The search can be **filtered out** in categories such as document id, text source, genre, gender, dialect.
- A **word list** is available to search for specific word forms.
- A **Cyrillic keyboard** was created to support typing.

Future plans

We plan to develop our corpus in the following areas:

- collecting, processing and including further Tundra Nenets texts
- creating a parallel Tundra Nenets–Russian corpus with sentence-level alignment
- adding linguistic annotations at certain levels, e.g. tagging sentence-types
- creating CIMDI metadata for the texts

Acknowledgements

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.