A digital text collection of Tundra Nenets

Nikolett Mus
Hungarian Research Centre for Linguistics
mus.nikolett@gmail.com

Relevant UniDive working groups: WG1, WG3, WG4

1 Introduction

The paper reports on an ongoing corpus-building work of the Tundra Nenets language (Northern Samoyedic, Uralic). The work is carried out within the frame of the research project titled "Theoretical and experimental approaches to dialectal variation and contact-induced change: a case study of Tundra Nenets" (ID: NKFIH 129235).[1] The corpus-building methods and the work have been developed and undertaken by the Author and Réka Metzger who is a member of the project. I will describe here our methodology and discuss the main lessons drawn from our work.

There are Tundra Nenets corpora (or annotated witten and spoken texts) available on the web, nevertheless, these sources primarily serve to sample the language, and cannnot be considered as large, robust, balanced, and/or representative corpora. The language is endangered and the speaker community is not active in archiving (and/or revitalizing) their language. Although the grammatical profile of the language is of great interest for theoretical linguists, the language is not sufficiently supported digitally. An example of an unusual cross-linguistic phenomenon in Tundra Nenets is the grammatical property of Nominal predicates. Tundra Nenets licenses copula omission with nominal, adjectival, and numeral predicates in all numbers and persons in the present tense and in the past tense. The Nominal predicate bears a subject agreement suffix and a past tense marker (1).

(1) *măń lekara-dam-ź.*
1SG doctor-1SG-PST
'I was a/the doctor.'

It is to be noted that the agreement on the Nominal is (intransitive) verbal agreement (2).

(2) *măń ma-kańi* *xăja-dm?.*
1SG tent-LOC.POSS.1SG go-1SG
'I go to my tent.'

[1]For further info please visit the webpage https://tundranenetsdata.nytud.hu/index.html

There is no vibrant research community examining the language. The lack of digital and human resources complicates teaching and researching the language. The aim of our work is to fill this gap and provide a tool that can be used for this purpose. Furthermore, our text collection may serve as the base of a language model or various language tools. Our method allowed us to compile and process a collection that contains 491,700 tokens in a relatively short period of time.

2 Background

Tundra Nenets is an endangered indigenous minority language spoken by c. 20,000 people in the North-Eastern part of Europe and in the North-Western part of Siberia in three administrative districts of the Russian Federation. A few more groups of speakers are sporadically found in the neighbouring areas. There is no (written or spoken) variety of Tundra Nenets that is standard for all speakers. Instead, three dialect groups – and within them further (sub)dialects – are differentiated. These groups are exposed to different external and internal influences, and little is known about the structural difference found among them. The EGIDS status of Tundra Nenets is 6b (*threatened*), i.e., it is still used in oral communication in everyday interactions within all generations. Thus, efficient speakers are found among the members of the youngest generation. But, there is a continuous decline in the number of speakers, and one can hardly find a Tundra Nenets speaker who is not bi- or multilingual (Trevilla, 2009).

As for the *digital presence* of Tundra Nenets. We find online newspapers, a test version of Wikipedia, videos and audio recordings (provided and archived by the Yamal Region broadcast) of Tundra Nenets.[2] There is also a dataset containing word and character n-gram frequencies for Tundra Nenets in IPA provided by the An Crúbadán

[2]http://nvinder.ru/rubric/yalumd https://incubator.wikimedia.org/wiki/Special:PrefixIndex/Wp/yrk/ http://yamal-region.tv

Project.[3] Finally, certain Tundra Nenets language tools, such as a text analyzer, paradigm and (number) word generators, a digital dictionary, are found on the website of Giellatekno.[4] Additionally, a relatively huge amount of printed texts representing various genres are published in print (sometimes together with Russian translations). As to our knowledge the speaker community is not/barely involved in research projects.

The only variety of Tundra Nenets that is systematically described so far is the one spoken in the Yamal Peninsula (belonging to the Eastern group). Reference grammars (Nikolaeva, 2014) and dictionaries (Tereshchenko, 1965) of this variety are found.

A small (mainly descriptivist) research community in Russia and in Europe deals with (Tundra) Nenets, but the research field is fragmented, and the members of the research community are largely disconnected.

3 Methodology and new results

As a starting point, we collected Tundra Nenets printed (written) materials and texts from the web. A significant amount of Tundra Nenets texts is available *in print*. After scanning these sources, .pdf files were converted into .txt ones by using an optical character recognition software.[5] OCR accuracy was checked manually by comparing the output files with the original texts. Given that there is no spell checker of Tundra Nenets, we applied some phonotactic rules to increases the OCR accuracy: we searched for character sequences that violate phonotactic rules of Tundra Nenets, e.g. Tundra Nenets does not allow initial consonant clusters. These were then manually corrected in the final output files. Newspaper articles from the *web* were saved automatically. We implemented a web scraper in Python that collects URLs of the articles, then iterates through them and extracts the necessary metadata and data from the HTML tags using regular expressions.

The next step was to *standardize* the raw data. The sources were converted into UTF-8 encoded .txt files, cleaned from extra white spaces, and the punctuation was unifyed. As mentioned, there is no written standard of Tundra Nenets. We faced two types of encoding problems specific to the language: (i) in certain texts the same character was used with different (grammatical) functions, and (ii) different characters/graphemes stood for the same phoneme. To find satisfactory solutions, we consulted with members of the speaker community.

Since different written representations of words might reflect dialectal differences we did not aim at unifying and/or normalizing the words. Thus, word forms did not go through any further modification (although a normalization process will be necessary).

Finally, we prepared the texts for the corpus management system. We use the open-source version of Sketch Engine *corpus management system*, i.e. NoSketch Engine (NoSkE) (Kilgarriff et al., 2014). NoSkE offers great searching features already at this level of text processing, i.e. it allows both simple searches, for instance, occurrences of characters, words, word forms, and more complex queries by using regular expressions. In addition, our broader goal is to create a Tundra Nenets–Russian parallel corpus. NoSkE is able to store texts from various languages, and the parallel concordance is not affected by the difference in the level of analysis of the parallel texts. NoSkE requires two files to compile a corpus. Each text has to be converted into an XML format vertical file, where every token and its metadata, e.g. lemma, POS-tag, are in a separate line. It is possible to define attributes in the root tag that allows the user to filter the search by these categories. We merged the XML files into one vertical file as an input of NoSkE. In addition, a corpus configuration file is needed. This defines the structure of the corpus and contains additional information, e.g. language, encoding, description, etc.

We modified the user interface of NoSkE by customizing the menu, and created a Cyrillic keyboard to make the corpus search easier.[6]

In order to enable a *sampling frame* we catalogized the dialect (group)s, gender, and age of the speakers/informants, and the date of recording. Furthermore, we classified the sources by the reality of the speech event when the texts were constituted (Schneider, 2002). We differentiated written texts that were composed on real speech situations, i.e. narrative folklore texts, phrasebooks

[3] http://crubadan.org/languages/yrk-x-tundra-acad

[4] http://giellatekno.uit.no/cgi/index.yrk.eng.html

[5] Note that Tundra Nenets writing system is based on the Cyrillic alphabet.

[6] The corpus is available here https://tundranenetsdata.nytud.hu/bonito

and methodological handbooks, from texts that were produced in imagined situations and have never been spoken, i.e. newspaper articles.[7] This procedure was carried out manually. In the current form of the collection, one can divide texts by these data.[8] It is to be noted that our text collection does not meet the expectations of a *reliable*, *natural*, *balanced* and *representative* collection (Himmelmann, 1998; McEnery and Hardie, 2011).

4 Future plans and possibilities

We plan the following steps to develop our data set:

- to automatize some of our processes, e.g. OCR;
- to normalize texts;
- to create a sentence-level aligned parallel Tundra Nenets–Russian corpus;
- to contact researcher and speaker communities for customizing the corpus and the user interface.

References

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36:161–196.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Tony McEnery and Andrew Hardie. 2011. *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Irina Nikolaeva. 2014. *A grammar of Tundra Nenets*. De Gruyter Mouton.

Edgar W Schneider. 2002. Investigating variation and change in written documents. *The handbook of language variation and change*, 67:96.

Natalia M. Tereshchenko. 1965. *Nenetsko-Ruskii slovar [Nenets-Russian dictionary]*. St. Petersburg.

Lorena Trevilla, editor. 2009. *Ethnologue: Languages of the World*. SIL International.

[7] In the Uralic fieldwork tradition, it was/is an established practice to collect and record narrative folk stories during ethnographical and linguitsic fieldworks. These texts were/are transcribed and edited for publishing in books/collections. Many of the original recordings/transcriptions are not accessible, therefore the extent of the editing procedure is not salient. So these texts are not transcriptions of spoken data.

[8] We created a catalogue of these metadata. The catalogue can be accessed on the following website https://tundranenetsdata.nytud.hu/tundra_nenets_materials.html