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

Nikolett Mus
Hungarian Research Institute
for Linguistics
musn@nytud.hu

Réka Metzger
Hungarian Research Institute
for Linguistics
metzger.reka@gmail.com

Abstract

In this paper, we report on the main lessons drawn from the first year of a Tundra Nenets (Samoyedic, Uralic) corpus building work carried out in the Hungarian Research Institute for Linguistics. The aim of our work is twofold. First we collect, process and archive written (and in the latter part of the project period spoken) data of Tundra Nenets. Second, we build a parallel corpus, i.e. a Tundra Nenets–Russian corpus, to support and encourage preferably synchronic syntactic research on Tundra Nenets. After discussing certain language and culture specific factors that potentially influence the sampling method, we present the stages of our work in detail.

1 Introduction

When we started our research project titled “Theoretical and experimental approaches to dialectal variation and contact-induced change: a case study of Tundra Nenets” (ID: FK_129235) in 2018, the need arose to archive (i) the published and carefully selected written Tundra Nenets texts, which we collected to explore our preliminary syntactic research questions, and to formulate our working hypotheses within the frame of the project, and (ii) the transcribed spoken Tundra Nenets data, which were/will be collected during our fieldworks in the project period (between 2018–2022).[1] Even though there are Tundra Nenets corpora (or annotated witten and spoken texts) available from the web, these sources primarily serve to sample the language, but cannot be considered as large, robust, balanced, or representative corpora.[2] It was therefore concluded that building a Tundra Nenets corpus by processing our archived materials would fill the gap in the currently available resources.[3] We considered important the *reliability*, *naturalness*, *balancing* and *representativeness* of the various sampling ideals usually emphasized in the literature (Himmelmann 1998; McEnery and Hardie 2011). Our intention is to design a corpus that meets these requirements. Nevertheless, there is (at least) one aspect specific to Tundra Nenets that influences this goal. Certain characteristics of Tundra Nenets written and spoken sources cannot be validated and controlled for. For instance, the representativity of text varieties associated with speakers of different genders and age might not be balanced in the future corpus. Before we discuss the stages of our work, let us provide some background information on the language itself.

Tundra Nenets (Samoyedic, Uralic) is one of the many endangered indigenous minority languages of the Russian Federation. Its status is 6b, i.e. *threatened*, on the EGIDS scale. It means that Tundra Nenets is still used in oral communication in everyday interactions within all generations. Thus, efficient speakers are found among the members of the youngest generation. However, there is a continuous decline in the number of speakers (Trevilla, 2009). The language is spoken by c. 20,000 people in the North-

[1] For more details and information of the research project, please visit our website: http://www.nytud.hu/oszt/elmnyelv/thea/index.html

[2] Tundra Nenets audio recordings are found on the project website “Endangered Languages and Cultures of Siberia” (http://www.siberianlanguages.surrey.ac.uk/about/), as well as, in ELAR (https://elar.soas.ac.uk/Collection/MPI120925). The projects titled “Typology of Negation in Ob-Ugric and Samoyedic Languages (NOS)” (https://www.univie.ac.at/negation/sprachen/nenetsa.html), and “Languages under the Influence. Uralic syntax changing in an asymmetrical contact situation” (http://www.nytud.hu/depts/tlp/uralic/dbases_tundranenets.html) provide morphologically annotated Tundra Nenets texts.

[3] The corpus is available on the following website https://tundranenetsdata.nytud.hu/bonito

Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages: Vol. 2 Resource Papers and Extended Abstracts, pages 4–9,
Online, March 2–3, 2021.

Eastern part of Europe and in the North-Western part of Siberia in three major administrative districts of the Russian Federation. These are the Nenets Autonomous Okrug, the Yamalo-Nenets Autonomous Okrug and the Taymyrsky Dolgano-Nenetsky District. Additionally, 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. The language itself consists of three main dialectal groups. Within them, one can distinguish further (sub)dialects. Little is known about the difference found among these dialectal groups and dialects.[4] The speakers live on a relatively large territory together with other indigenous minorities in the area: there are Khanty (Uralic), Mansi (Uralic) and Selkup (Uralic) speakers in the Yamalo-Nenets Autonomous Okrug, Nganasans (Uralic), Tundra and Forest Enets (Uralic), Dolgans (Turkic), Kets (Yeniseic) and Evenkis (Tungusic) in the Taymyrsky Dolgano-Nenetsky District, and Komi speakers (Uralic) in the Nenets Autonomous Okrug. Additionally, the Russian language and culture has a great influence on the Tundra Nenets speaking community. Thus, the language is exposed to different external and internal influences, and one can hardly find a Nenets speaker who is not bi- or multilingual. There is neither a unified literary language, nor a unified writing system of Tundra Nenets. This is partly due to the fact that the (Tundra) Nenets literacy does not have a long history and tradition.[5]

As for the digital presence of Tundra Nenets, there are online Tundra Nenets newspapers, e.g. Няръяна вындер ('Red Tundra').[6] In addition, a test version of Tundra Nenets Wikipedia is also present.[7] Tundra Nenets videos and recordings are provided and archived by the Yamal Region broadcast.[8] There is also a dataset, that contains word and character n-gram frequencies for Tundra Nenets in IPA provided by the An Crúbadán Project.[9] Tundra Nenets language tools, for instance a text analyzer, paradigm and (number) word generators, and digital dictionaries, are available on the website of Giellatekno.[10]

2 Our methods and results

The workflow in Figure 1 shows the individual stages of our corpus-building work. Green colour marks those stages that were automated, i.e. by using computational methods. The blue colour indicates the works/tasks carried out manually.

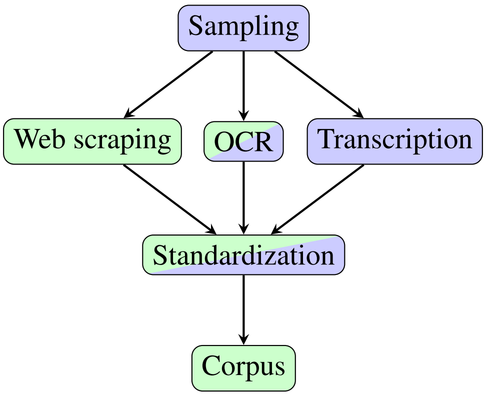

Figure 1: Workflow of corpus building process

2.1 First stage: sampling

In the **first stage**, we determined the **sampling frame** of our corpus. Our intention was to equally represent the *dialect(al group)s* of the language, the *gender* and *age* of the speakers, and the *date* of the recording. Even though our corpus does not meet any of these expectations in its current stage, our aim is to supplement our data with the missing ones, for instance, by collecting materials during fieldworks. Therefore, we created a catalogue, which we can use as a base during our future text collecting procedure.[11] An additional criteria was to represent both *spoken* and *written* varieties of the language. This type of differentiation of text types was particularly important as preliminary syntactic studies indicate that there are grammatical differences between the spoken language and its written form (Asztalos et al., 2017). The spoken texts, whose amount in

[4] The very purpose of our research project is to compare the syntax and prosody of certain constructions in two of these Tundra Nenets (sub)dialetcs.

[5] The intention (originated from external needs) to create unified literary languages and writing systems of the indigenous people of Western Siberia arose only in the late 1920s and early 1930s (Toulouze, 1999). This resulted in a writing system of Tundra Nenets based on the Cyrillic alphabet.

[6] http://nvinder.ru/rubric/yalumd

[7] https://incubator.wikimedia.org/wiki/Special:PrefixIndex/Wp/yrk/

[8] http://yamal-region.tv

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

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

[11] The catalogue is available on our website https://tundranenetsdata.nytud.hu/index.html#corpus.

the current corpus is very limited (1246 tokens), were recorded by us during consultations with a native speaker, and transcribed by the speaker himself. Following Schneider (2002), we classified our written sources on the basis of the reality of the speech event when the texts were created. On the basis of this classification, we differentiated written texts that are potentially based on a real situation in a given time and place from those that represent texts produced in imagined situations, i.e. they have never been spoken. For instance, folklore texts belong to the former group. In the Uralic fieldwork tradition, it was/is an established practice to collect and record narrative folk stories during ethnographical and linguitsic fieldworks. Then, these texts were/are transcribed and edited for publishing in books/collections. Unfortunately, many of the original recordings/transcriptions are not accessible, therefore we are not aware of the extent of the editing procedure. This is the reason of not classifying these texts as real transcriptions of spoken texts. However, we do not consider them to be equal to the latter type of texts either. Outside of folklore texts, of which we only kept the narrative ones, phrasebooks and methodological handbooks belong to this group. Texts representing the written variety of the language are newspaper articles that were originally created in writing. We included issues from two newspapers: Няръяна вындер ('Red Tundra') and Няръяна ӈэрм ('Red North'). The former is available from the web, and we included the Tundra Nenets articles between 14/02/2013–18/04/2019. Articles of the latter newspaper were selected randomly, because we have limited acces to the newspaper itself, i.e. it is available in print.

In the current form of the Tundra Nenets corpus, one can divide the corpus by the *genre* and the *source* of the texts, and by the *age* and the *dialect* of the speakers. Note, that in some sources we were not able to extract all of the required metadata. As it is indicated in Figure 1, the sampling procedure has completely been carried out manually.

2.2 Second stage: text processing

The **second stage** of our work was the **process** of converting raw data into UTF-8 encoded .txt files. The idea of creating a digital corpus required to have all written data in the same format. There were three types of texts we had to process in this stage: (i) written texts available from the web; (ii) (published) written sources in print; (iii) recorded spoken texts. In order to convert the texts into machine-readable form, the texts had to be (pre-)processed in different ways depending on their characteristics:

- texts in (i) ⇒ web scraping
- texts in (ii) ⇒ (scanning and) OCR
- texts in (iii) ⇒ transcription

2.2.1 Web scraping

Saving hundreds of articles from the web manually would have taken a very long time. To make the harvesting process of the online newspaper less time consuming we implemented a **web scraper** in Python. It collects URLs of the articles, then iterates through them and extracts the necessary metadata and data from the HTML tags using regular expressions. The output was 793 Tundra Nenets plain texts.

2.2.2 OCR

A significant amount of Tundra Nenets texts were available in print. After scanning these sources, we converted the .pdf files into .txt ones by using the **optical character recognition** software developed by Abbyy FineReader.[12] Then, we checked the accuracy of OCR by manually comparing the input and output files. With the aim of simplifying and accelerating the procedure in the second process for checking the output files, we relied on some phonotactic constraints of Tundra Nenets.[13] We listed the words of the OCRed texts and searched for forms that violates these phonotactic rules. These were then manually corrected in the final output files. The output of this process was 243 Tundra Nenets texts.

2.2.3 Transcription

We recorded 5 Tundra Nenets texts in 2017.[14] These texts were **transcribed** by our Tundra Nenets informant. The transcription of these texts is a simple orthographic transcription.

[12] https://pdf.abbyy.com/

[13] In Tundra nenets, (i) vowels, (ii) consonant clusters, and (iii) voiced/weak/lax consonants, velar and glottal stops, affricates, and trills cannot occur in word initial position.

[14] The audio files are available on the website http://corpus.nytud.hu/people/eszter/uralic_preerc/yrk/uj_szovegek/Okotetto/audio/.

2.3 Third stage: standardization

After converting the data into the same format, a **standardization** process started as the **third stage** of our work, i.e. cleaning the texts from extra white spaces and unifying certain characters. In order to make the corpus easily searchable, we unified the punctuation.

After this, we had to find a solution for two encoding problems specific to Tundra Nenets:

1. the same character was used with different (grammatical) functions,
2. different characters/graphemes stood for the same phoneme.

A typical example of (1) above is that the standard double quotation mark (U+0022) is used in quotations, and it also stands for a glottal stop phoneme in the texts. For distinguishing glottal stops from the actual quotation marks, we inserted the quotes between French quotation marks (U+00AB, U+00BB).[15] Since there was no difference between the glottal stops and the ending quotation marks indicated in the texts, this replacement process was not fully automated: all the ending quotation marks were listed and manually checked in the texts.

As regards (2), we found, for instance, that both apostrophe (U+0027) and standard double quotation mark (U+0022) were used for the glottal stop phoneme. There are historical reasons for differentiating two glottal stops in the language, however, the underlying forms cannot be regarded as two phonemes synchronically.[16] Despite we adopt the existence of only one glottal stop phoneme in the language, we kept the original differentiation in the texts. This is because consultations with the speaker community revealed that the vast majority of speakers themselves differentiates two glottal stops in writing.[17]

Another example of (2) is the representation of the velar nasal phoneme in the texts: three different graphemes were used (U+04C9, U+04A2, U+04C8). Since there is no linguistic explanation of using three different graphemes, we unified them by using only one in the texts. After consulting with the community, we kept the one (U+04C8) that is standardly used in the virtual keyboard apps, e.g. in Gboard. Table 1 shows the distribution of these characters in the original texts grouped by genre.[18] Texts from newspapers represent the current language use. The numbers also support that the grapheme we use in our corpus is the preferred one in contemporary texts.

Genre	U+04A2	U+04C8
Folklore	29653	247
Methodological hb.	484	65
Narratives	225	0
Newspapers	391	47208
Phrasebook	351	0
Sum	31104	47520

Table 1: Distribution of velar nasal graphemes

These decisions were made in support of making the corpus searchable while not losing any additional information that can be essential for linguistic research. As mentioned in Section 1, there is no written standard of Tundra Nenets. Since different written representations of words might reflect dialectal differences, as it is in the case of, e.g. тикы 'that' (Yamal) and чикы 'that' (Taymyr), we did not aim at unifying and normalizing the texts. Thus, word forms did not go through any further modification.

2.4 Fourth stage: corpus

Preparation of our data for **NoSketch Engine** (NoSkE) corpus management system was the **fourth stage** of our work. We use the open-source version of Sketch Engine (Kilgarriff et al., 2014). In selecting the most appropriate corpus query system, we considered the following relevant factors. (i) Our texts have not yet been analyzed and/or annotated. Therefore, we needed a management system that is powerful enough to offer great searching features already at this level of text processing. NoSkE fulfils this criteria as it allows both simple searches, e.g. occurrences of characters, words, word forms, and more complex queries by using regular expressions. In the latter case, one can define search patterns in order to get more pre-

[15] Note that glottal stops in Tundra Nenets mark the plural and the genitive forms of nouns, and play role in the verbal conjugation paradigm, a.o.

[16] As Hajdú (1968) and Staroverov (2006) note, the glottal stops are pronounced in the same way and do not differ in any acoustic properties.

[17] Interestingly, our corpus data shows that the use of the two graphemes is not consistent at all. Thus, this marking strategy is accepted as a norm in the community but does not have a real function.

[18] One of the graphemes (U+04C9) occurred only once therefore we do not represent it in Table 1.

cise results. (ii) Our broader goal is to create a Tundra Nenets–Russian parallel corpus. We expected from a management system the ability to store texts from various languages. In NoSkE, furthermore, the parallel concordance is not affected by the difference in the level of analysis of the two texts. (iii) Tundra Nenets is not sufficiently supported digitally. Therefore, we wanted to have a system that allows the possibility to create a corpus with such an unsupported language. (iv) Having a clear, user-friendly and customizable interface was also on our list. This ensures, that our corpus can/will be used by non-linguists, too.

NoSkE requires two files to be able to compile a corpus. First, each text has to be converted into an XML format file that contains the texts vertically. In these vertical files, every token and its metadata, e.g. lemma, POS-tag, are in a separate line.[19] Tags are also used to structure the data, e.g. to divide the text into paragraphs or sentences, or to mark when there should not be space between a word and the following punctuation. It is also possible to define attributes in the root tag that allows the user to filter the search by these categories. In the current form of our corpus, the *id*, *source*, *genre*, *gender* and *dialect* attributes of the texts are defined. After converting the raw .txt files into the expected file format automatically, we merged the XML files into one large vertical file as an input of NoSkE. Second, it is also necessary to add a corpus configuration file. The corpus configuration file defines the structure of the corpus and contains additional information about it, such as language, encoding, description etc. After creating both files, NoSkE was able to successfully compile our Tundra Nenets (Monolingual) corpus with its 452,930 tokens.[20]

As mentioned in Section 2.1, the following metadata of the texts were collected: *source*, *date* of origin, *type* (written/spoken), (*sub*)*genre*, *number of tokens* and the *name*, *gender*, *age* and *dialect* of the speaker. We stored these information in an .xlsx catalogue.[21] This format can later be converted into XML files required by the IMDI/CMDI metadata framework, which we aim at using in our project. This systematic way of storing the metadata helps us to take into account the weaknesses of our corpus/archive.

We also modified the user interface of NoSkE by customizing the menu: we retained those features that can be useful at this level of text processing/annotation.

Given that our corpus contains only Tundra Nenets texts in the moment, it is possible to search Tundra Nenets data, such as phonemes, phoneme clusters, morphemes, words, etc. in it. To make the search easier, we created a Cyrillic keyboard.

3 Ongoing work and future plans

In the first year of our corpus building work, we set up the framework of the corpus building process and automatized these stages as far as possible. The development of our corpus is found to be feasible in the five main area highlighted below:

- to collect, process and include more texts. To do so, we plan to contact archives and individual researchers, besides we participate in fieldworks. Our focus is on balancing the representation of dialects, and written/spoken texts;
- to automatize some of the stages of our process described above in Section 2;
- to create a parallel Tundra Nenets–Russian corpus: the Russian translations of our texts will be aligned at sentence level as the first step;[22]
- to annotate the Tundra Nenets data at certain levels, e.g. tagging questions;
- to contact researcher and speaker community in order to customize both the corpus and the user interface to their needs.

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.

[19] Given that our texts do not have any annotation yet, there are only tokens in each line so far.

[20] https://tundranenetsdata.nytud.hu/bonito

[21] It is available on our website: https://tundranenetsdata.nytud.hu/index.html#corpus

[22] Some of the translations are already available. The other texts will be translated by Tundra Nenets–Russian bilingual speakers.

References

Erika Asztalos, Katalin Gugán, and Nikolett Mus. 2017. Uráli vx szórend: nyenyec, hanti és udmurt mondatszerkezeti változatok. In É. Kiss Katalin, Hegedűs Attila, and Pintér Lilla, editors, *Nyelvelmélet és diakrónia 3*, pages 30–62. PPKE BTK, Budapest.

Péter Hajdú. 1968. *Chrestomathia Samoiedica*. Tankönyvkiadó.

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.

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

Peter Staroverov. 2006. Vowel deletion and stress in tundra nenets. In *Proceedings of the first central European student conference in linguistics*, pages 1–20. Research Institute for Linguistics, Hungarian Academy of Sciences.

Eva Toulouze. 1999. The development of a written culture by the indigenous peoples of western siberia. *Arctic Studies*, 2:53–85.

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