Starting a UD Treebank of Tundra Nenets

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

Background

Tundra Nenets (Samoyedic, Uralic) **written and spoken texts** may be found in a number of archives, including the ELAR archive, however the **annotation** of these sources is **not consistent**, if it exists at all. For some languages in the Finno-Ugric branch of the Uralic language family, **UD Treebanks** are available; however, **none** exist **for the Samoyedic branch**.

⇒ The aim of my project is to start a new UD Treebank for Tundra Nenets.

The Tundra Nenets language

Tundra Nenets (Samoyedic, Uralic; ISO 639-3 code: yrk; extended language code: yrk-tun) is an endangered indigenous languages spoken in the Russian Federation. 20.000 Tundra Nenets speaks the language (as L1). The language has an EGIDS classification of 6b, which is *threatened*. The culture of Tundra Nenets is predominantly an oral one without a unified literary language, and/or a unified writing system. The language has been influenced by the Russian language as well as other indigenous minorities in the region. Tundra Nenets is an agglutinative-concatenating and left-branching language whose digital support is not sufficient.

The dataset of the project

The **starting materials** of the Tundra Nenets UD Treebank are (mainly unpublished) texts, i.e. transcripts of spoken texts, that were gathered from a **native speaker informant** during consultations in Moscow in 2017. The informant speaks the Yamal dialect of Tundra Nenets. (The speaker has given a **permission** to publish this data.) During the data-collection period, methods adhered to the **standard protocols** of modern **linguistic fieldwork** were employed, i.e. **semi-controlled natural language production data** was obtained using interactive, goal-driven, real-time conversational activities.

Task type	Genre	Tokens	Sentences
Route description 1–3	narrative	489	91
”Arctic reindeer”	narrative	300	53
”Pear Story”	narrative	456	77
Total		1,245	221

Table 1: The metadata of the texts

⇒ The Treebank will represent a **contemporary**, non-edited, **colloquial** variation of **Tundra Nenets**.

Methods and objectives of the project

Completed tasks

Transcription

The conversations were recorded during the sessions, and the informant later (**orthographically**) **transcribed** these recordings using the Tundra Nenets alphabet, which is an extended version of the **Cyrillic script**.

(Character) standardisation/unification

The transcripts were standardised/unified, which mainly involved **unifying** certain **characters**.Two types of problems were solved:

- the same character was used for different functions;
- different characters were used for the same function.

Translation and sentence-level alignment

Russian translations of the transcribed texts were done by the informant. In addition, the Tundra Nenets texts were translated into English. The parallel texts in **Tundra Nenets**, **Russian**, and **English** were manually **aligned at the sentence level**.

Morphological analysis and POS tags

The morphological analysis of the data collection was completed manually: **POS tags** and **morphological labels** were added to the words. The Tundra Nenets-specific enhanced version of the **L**eipzig **G**lossing **R**ules were applied during the analysis.

Open issues

Normalisation

One of the biggest challenge of this project is to solve the problems of the **orthographic variation** exhibited in the written texts of Tundra Nenets. The language does not have a standard spelling. Thus, selecting and setting a baseline are causing the major difficulties.

Q: What rules and methods for normalising words of a language lacking a standard variation should be followed?
Q: How to decide a reference source that will be handled as the baseline?

UD treebank

Tundra Nenets is currently **not listed** on the UD website.

Q: Which tools related to Universal Dependencies should be used?
Q: Is it possible to (semi-)automatise the process of creating UD Treebanks?

Future plans

The goal is to extend the Tundra Nenets UD Treebank with additional sentences. Published newspaper texts comprising of around 330,000 tokens are prepocessed.