Title: Starting a UD Treebank of Tundra Nenets

Applicant's name and affiliation: Nikolett Mus, Research fellow in Hungarian Research Centre for Linguistics, Budapest, Hungary

Keywords: Tundra Nenets (Samoyedic), UD Treebank, fieldwork material, normalisation

1 Aims Even though there are Tundra Nenets (Samoyedic, Uralic) corpora (or written and spoken texts) available from the web, the annotation of these sources is inconsistent (if there is any). The currently accessible UD Treebanks represent the Finno-Ugric branch of the Uralic language family, while Treebank(s) of languages of the Samoyedic branch are not available. For these reasons, the aim of my project is to start a UD Treebank for the Tundra Nenets language.

2 Background Tundra Nenets (Samoyedic, Uralic; ISO 639-3 code: yrk; extended language code: yrk-tun) is one of the numerous endangered indigenous languages spoken in the Russian Federation. Approximately 20.000 Tundra Nenets speakes the language (as L1). The language has an EGIDS classification of 6b, which is *threatened* (Trevilla 2009). 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 insufficient; to the best of my knowledge Giellatekno and the Crúbadán Project offer some early-stage tools.[1]

3 Dataset of the project In Samoyedic linguistics, folklore texts of speakers of the oldest/older generation are traditionally (and typically) collected during linguistic fieldworks. The published and edited versions of these collections comprise the great majority of the available Tundra Nenets (written) sources. In order to avoid potential copyright issues regarding these published data and to present a contemporary, non-edited, colloquial variation of Tundra Nenets, I propose creating a UD Treebank using (mainly unpublished) texts – or more precisely transcripts of spoken texts – that I collected from a native speaker informant during our 2017 consultations in Moscow. The speaker has given me permission to publish this data. During the data-collection period, I employed methods that adhered to the standard protocols of modern linguistic fieldwork: semi-controlled natural language production data was obtained using interactive, goal-driven, real-time conversational activities. (The metadata of the texts I intend to analyze for the project are listed in Table 1 in Section 5.) Although I am aware of the relatively limited amount of data, I still regard this proposal as a pilot research, the results of which may constitute the theoretical and practical base of a future corpus buidling work.

4 Methods and individual objectives of the project This Section describes the steps and methods of the project.

4.1 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).

4.2 (Character) standardisation/unification Since the transcriptions are now a part of our Tundra Nenets Monolingual Corpus, we standardised/unified the texts, which mainly involved unifying certain characters (see Mus & Metzger 2021 for a detailed description of this process).[2]

4.3 Normalisation As was mentioned previously, Tundra Nenets lacks a unified spelling system, therefore there are (spelling) variations in the texts. Despite the small size of my current data set and the fact that it only comes from one speaker, this issue needs to be resolved at this point in the project to prevent conflicts in the data when I add more texts to the dataset in the future. This is the phase of my project for which I would require assistance in the methodological considerations.

4.4 Translation and sentence-level alignment The transcribed texts are accessible in both Russian and English (my translations are in English, while the informant contributed the Russian

[1] http://www.language-archives.org/item/oai:crubadan.org:yrk-x-tundra-acad and https://giellatekno.uit.no/cgi/index.yrk.eng.html

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

translations). The parallel texts in Tundra Nenets, Russian, and English were manually aligned at the sentence level.

4.5 Morphological analysis (and POS tags) The morphological analysis of a few texts from the data collection has already been completed. I applied the Leipzig Glossing Rules during the manual analysis and add POS tags and morphological labels to the words. The Tundra Nenets-specific enhanced version of the LGR tagset is accessible as a table in the.xlsx format. By the time of Summer School, I want to have finished all of the morphological analysis.

4.6 UD treebank Based on four criteria, I have selected languages for which UD treebanks are accessible. My criteria were 1) genetic affiliation, i.e. Uralic languages (e.g. Finnic branch, Mordvin branch, Permic branch, Sami branch, Hungarian); 2) (potential) contact languages, e.g. Slavic (Russian); 3) languages that may have similar areal features, i.e. Siberian languages (Yakut/Sakha); and 4) languages with typological similarities, i.e. agglutinative-concatenating, left-branching languages; e.g. Turkic, Eskimo-Aleut, Chukotko-Kamchatkan, Japanese languages.). I intend to read the guidelines of these Treebanks and use them as reference manuals in my project.

5 The current state of the project The project is in the process of creation, several steps have been undertaken. The current state of the process and a timeline are shown in Table 1 (green = done; orange = in production; red = assistance needed)

Task type	Tokens	Sentences	4.1	4.2	4.3	4.4	4.5	4.6
Route description 1–3[3]	489	91	🟩	🟩	🟥	🟩	🟩	🟥
"Arctic reindeer"[4]	300	53	🟩	🟩	🟥	🟩	🟧	🟥
"Pear Story"[5]	456	77	🟩	🟩	🟥	🟩	🟧	🟥
Total	**1.245**	**221**						

6 Explanation how the participation in the training school will be useful for the project One of the biggest challenge of this project is to solve the problems of the orthographic variation exhibited by the written texts of Tundra Nenets. I am not aware of the methods/methodological considerations of **normalisation** (e.g. which model should I use). Besides, Tundra Nenets does not have a standard spelling. Thus, selecting and setting a baseline are causing the major difficulties, i.e. how to decide the reference source that will be handeld as the baseline. It is my goal that attending the Summer School shall promote solutions to this problem. Besides, I don't have any experience **building Treebanks**; my primary area of expertise is descriptive and formal syntax. Advice and guidance on things like which annotator to use and how to set up the infrastructure would therefore be beneficial in the beginning.

7 Open questions

(1) rules and considerations of **normalisations** of texts lacking a standard variation

(2) possibilities and conditions of (semi-)**automatising** UD Treebank creation

(3) handling "**exceptional**" cases, such as i) covert morphemes in NP ellipsis, in subject and object pro-drop constructions, copula drop in nonverbal predication (subject agreement and tense morphology on Nominal predicates), juxtaposed coordination (of NPs/CPs) ii) possessive marked postpositions (*нер-ни'* in.front.of-gen.1sg 'in fron of me') and possessive marked postpositions as parts of the pronominal paradigm (*ня-на* 1sg-dat 'from me') iii) non-possessive use of possessive suffixes, iv) non-nominative subjects / non-accusative objects v) agreement in number with lexical topical objects vi) noun incorporation (*мя-тан*-тю- tent-dat-enter 'enter a/the tent') vii) expletive negation viii) non-finite subordination (ptcp, cvb, nlz): subject agreement via possessive suffixes; case marker/postposition in adjunct clauses (*Маханякубта*-' *ми-ма-ни' сер'* right.side-gen go-ipfv.an-gen.1sg during 'while I am walking on the right side')

[3] This is an adaptation of the original HCRC map task (https://groups.inf.ed.ac.uk/maptask), designed for a single speaker. In this scenario, I, the fieldworker, was the Instruction Follower, and the speaker was the Instruction Giver. The map was simplified and the labeled features were made more appropriate for the cultural context.

[4] The stimulus was a short movie.

[5] https://www.linguistics.ucsb.edu/research/pear-film

References

Mus, Nikolett & Metzger, Réka 2021. Toward a Corpus of Tundra Nenets: Stages and Challenges in Building a Corpus. In *Proceedings of the 4th Workshop on Computational Methods for Endangered Languages* Vol.2 (Resource Papers and Extended Abstracts), 2021.

Trevilla, Lorena (ed.) 2009. *Ethnologue: Languages of the World.* SIL International.