A digital text collection of Tundra Nenets

Nikolett Mus
Hungarian Research Centre for Linguistics, Budapest
mus.nikolett@gmail.com|nytud.hu

WG1, WG3, WG4

Introduction

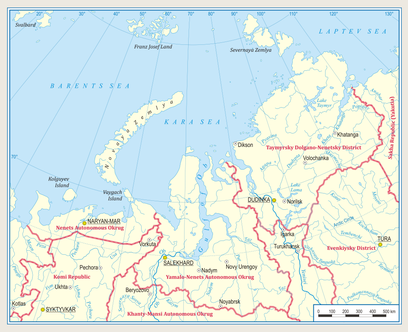
- an ongoing corpus-building work of the Tundra Nenets language
- a joint work with **Réka Metzger** (metzger.reka@gmail.com)
- https://tundranenetsdata.nytud.hu

Tundra Nenets

Population & family

- indigenous minority language
- spoken by c. 20,000 people
- belongs to the Samoyedic branch of the Uralic language family

Geography



Language vitality

- EGIDS status is 6b (*threatened*)
  - still used in oral communication in everyday interactions within all generations
  - there is a continuous decline in the number of speakers (Trevilla, 2009)
- no (written or spoken) standard
- three dialect groups (within them further (sub)dialects)
- only one variety is systematically described so far (Yamal < Eastern)
- the speaker community is not active in archiving (and/or revitalizing) their language

Digital presence & language support

- writing system based on the Cyrillic alphabet (Latin transcriptions)
- annotated written (and spoken) Tundra Nenets texts available on the web
- online newspapers, a test version of Wikipedia, as well as, videos and audio recordings provided and archived by the Yamal Region broadcast are found online

  ⇒ These sources serve to sample the language.
- a (relatively) huge amount of printed texts representing various genres are published in print (sometimes together with Russian translations)

  ⇒ There is not any large, robust, balanced, and/or representative corpus of Tundra Nenets.
- a dataset containing word and character n-gram frequencies for Tundra Nenets in IPA provided by the An Crúbadán Project
- text analyzer, paradigm and (number) word generators, a digital dictionary offered by Giellatekno
- online dictionary created within the frame of NOS project (in Vienna)

Language profile

- an agglutinative-concatenating language, e.g. *ŋăno-xona-ńi* boat-LOC-1SG 'in my boat' (with some fusional processes, e.g. *xaĺa* 'fish': *xali* 'fish:ACC.PL') (7 cases, 3 numbers, additional set in possessive paradigm)
- an accusative language (but! sometimes S marked with a local case, and O unmarked)
- obligatory subject agreement (person-number marking) and topical/given object agreement (of a 3rd person object) on transitive verbs, plus a reflexive paradigm (altogether 36 forms of verbal agreement suffixes)
- Nominal predicates bear a subject agreement suffix and a past tense marker (copula omission), e.g. *măń lekara-dam-ź* 1SG doctor-1SG-PST
- head-finality in APs, AdvPs, NPs, VPs (except in NegP *ńi-dm? mănes-?* NEG.AUX-1SG see-CONNEG)
- SOV (but! VX is also found)
- non-finite embedding (action nouns, participles, converbs) without complmenetizers (subordinate clauses usually precede the finite matrix clause)
- juxtaposed coordination and/or coordinate conjunctions

Acknowledgements

This study was funded by the National Research, Development and Innovation Office Hungary (NKFIH), grant FK 129235.

Methodology & results

Sampling (metadata)

- Tundra Nenets texts catalogized by "standard" metadata
  - dialect (group)s, gender, and age of the speakers/informants
  - date of recording/processing
- additional classification by the "reality" of the speech event (Schneider, 2002)
  - texts composed on real speech situations, e.g. narrative folklore texts, phrasebooks, methodological handbooks
  - texts produced in imagined situations (have never been spoken) e.g. newspaper articles

⇒In the current form of the collection, one can divide texts by these data.

Pre-processing

- printed (written) materials AND texts from the web selected/collected

Printed texts → OCR

- ABBYY FineReader
- accuracy checked manually
- double check by applying phonotactic rules (searched for character sequences violating phonotactic rules, e.g. Tundra Nenets does not allow initial consonant clusters)

Online texts → web scraping

- a web scraper in Python
- collecting URLs, iterating through them, extracting metadata and data from HTML tags using regular expressions

⇒ output: UTF-8 encoded .txt files

Standardizing

- raw data cleaned from extra white spaces
- punctuation unified
- two types of encoding problems specific to the language

Same character–Different function

- standard double quotation mark (U+0022) used in quotations AND stands for a glottal stop phoneme

⇒ The phonemic function has been kept.

Same function–Different character

- three different graphemes (U+04C9, U+04A2, U+04C8) mark the velar nasal phoneme

⇒ The one used in virtual keyboard apps, e.g. in Gboard has been kept.

⇒ The text collection is searchable (but word forms are not normalized).

NoSketch Engine (NoSkE)

- open-source version of Sketch Engine corpus management system (Kilgarriff et al., 2014)
  - texts converted into an XML format vertical file (tokens and metadata in separate lines)
  - XML files merged into one vertical file
  - corpus configuration file defining the structure of the corpus
- user interface of NoSkE modified by customizing the menu
- Cyrillic keyboard created

⇒ https://tundranenetsdata.nytud.hu/bonito

Future plans

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

References

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. 2014. The sketch engine: ten years on. *Lexicography* 1(1):7–36. · Schneider, E. W. 2002. Investigating variation and change in written documents. *The handbook of language variation and change*, 67:96. · Trevilla, L. (ed.) 2009. *Ethnologue: Languages of the World*. SIL International