From spoken data to UD treebank: a semi-automated methodology for the low-resource endangered language Tundra Nenets and beyond

Nikolett Mus
ELTE NyTK

Morgane Bona
Paris Nanterre University

Bruno Guillaume
INRIA

Aleksandra Miletić
Paris Nanterre University

Sylvain Kahane
Paris Nanterre University

Daniel Zeman
Charles University

Relevant UniDive working groups: WG1, WG4
Track: Work-in-progress

1 Aim and scope

This paper presents the development of a standardised, semi-automated workflow for processing spoken data and creating a richly annotated Universal Dependencies (UD) treebank for Tundra Nenets, an endangered and severely under-resourced Uralic language. Previous NLP-related work on Tundra Nenets has been sporadic and methodologically heterogeneous, lacking shared standards and interoperable annotation practices. The work was carried out within the framework of the UniDive COST Action, which enabled the design of a unified pipeline linking all stages of processing, from raw audio to syntactically and discourse-annotated corpus. The workflow integrates AI-based tools, rule-based annotation, and manual linguistic analysis, providing a reproducible approach to treebank creation and a transferable methodology for low-resource languages. The study is based on c. 52 minutes of unpublished spoken data collected during fieldwork with a Tundra Nenets speaker in 2017. The dataset includes semi-controlled elicitation tasks, such as map tasks, narrative and video-based storytelling, picture-based sequences, and guided conversations, designed to capture naturalistic language use, as well as scripted dialogues for controlled syntactic and prosodic analysis. A subset of these data (170 sentences, c. 12 minutes) has already been processed within the current workflow.

2 Background

Tundra Nenets is an endangered Uralic language spoken in northern Russia, characterised by rich agglutinative morphology, complex morphosyntactic structure, and discourse-configurational properties (Nikolaeva, 2014; Burkova, 2022; Mus, 2023). Despite its typological significance, it remains severely under-resourced from a computational perspective. Although Tundra Nenets is relatively well described compared to other languages in the region, and some digitally available corpora exist (Budzisch et al., 2025; Nikolaeva and Garrett, 2015), these resources remain heterogeneous. They differ in script (Cyrillic vs. Latin), morphological tagsets and annotation schemas (including the labeling of morphemes), as well as in data formats, and are only partially searchable. Interoperable, syntactically annotated corpora aligned with widely adopted frameworks remain absent. From the perspective of natural language processing, these conditions pose significant challenges. The lack of standardised, multi-layer annotated corpora limits the applicability of existing tools and hinders cross-linguistic comparability. Although recent advances in AI and multilingual modeling promise broader language coverage, languages such as Tundra Nenets are typically not explicitly represented, resulting in substantial performance gaps. Moreover, Tundra Nenets can serve as a valuable test case, providing a methodological model for the digital development of other languages in the region. The creation of interoperable resources is therefore essential not only for computational applications but also for linguistic research, language documentation, and long-term digital preservation.

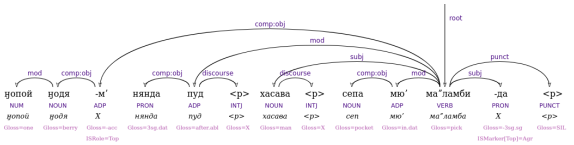
3 Methodology and Results

The workflow consists of a sequence of theoretically informed processing steps tailored to Tundra Nenets, combining AI-based tools, rule-based annotation, and manual analysis in a semi-automated pipeline. As mentioned, the project targets c. 52 minutes of unpublished spoken data, of which around 12 minutes have been processed so far, with further data planned to extend the annotation.

The first stage of the workflow employs state-of-the-art ASR models (Whisper3, Wav2Vec2, Omnilingual) to generate preliminary transcriptions. While performance is constrained by the extremely low-resource conditions, ASR output provides a useful starting point, reducing manual effort. The transcribed data are subsequently annotated following UD-based methodologies for spoken language (Kahane et al., 2021a), with adaptations specific to

Tundra Nenets. Annotation targets both lexical and non-lexical elements relevant for syntactic analysis, including pauses, disfluencies, false starts, and interruptions, ensuring accurate segmentation and interpretation while maintaining a reproducible, language-specific workflow.

Syntactic annotation is based on the mSUD framework (Kahane et al., 2021b; Guillaume et al., 2024), which captures morphology-syntax interactions central to the language, such as agreement and case marking. In addition, morphological information, including segmentation, is systematically incorporated into the annotation. While these steps are currently performed manually, ongoing work aims to extend the semi-automated pipeline to cover morphological analysis as well. Annotation is informed by existing corpora and grammatical descriptions of Tundra Nenets, and implemented semi-automatically using ArboratorGrew (see Figure 1 for an example). The outputs are subsequently converted into standard UD representations and validated for interoperability.



'The man slips the fruit one by one into his pocket.'

Figure 1: A Tundra Nenets example sentence annotated in mSUD

Given the discourse-configurational nature of Tundra Nenets (Nikolaeva, 2014; Burkova, 2022; Mus, 2023), an additional Information Structure layer (Topic, Focus, Contrast[1]) is envisaged. In particular, certain morphosyntactic patterns directly reflect discourse functions in the language; for example, object agreement suffix on the verb signals topicalised objects, encoding its discourse role independently of surface word order. While this layer has not yet been fully implemented, Tundra Nenets has served as a case study for the development of Information Structure annotation methodology within WG1 of the UniDive COST Action.

Finally, transliterations and translations are added to enhance accessibility. Complementary resources, including Latin-based transliterations and English translation, are provided to enhance accessibility and cross-linguistic applicability.

[1]In this framework, contrast is not treated as a fully independent category, but rather as a feature of Topic and Focus, i.e. contrastive topic and contrastive focus.

4 Conclusions and future plans

The workflow demonstrates how diverse, semi-controlled spoken data can be systematically processed for a severely under-resourced language such as Tundra Nenets. Developing the pipeline required decisions grounded in the language's structural and discourse properties, including the creation of semi-standard conventions for spoken-language phenomena and the integration of Information Structure role annotation. Future work will focus on extending the workflow to additional data, further standardising annotation practices, and incorporating new resources, aiming to create a more comprehensive and interoperable treebank for Tundra Nenets and other (low-resource) languages in the region.

Acknowledgements

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

References

Josefina Budzisch, Beáta Wagner-Nagy, and Alexandre Arkhipov. 2025. User's guide to inel nenets corpus.

Svetlana Burkova. 2022. Nenets. *Marianne Bakró-Nagy–Johanna Laakso–Elena Skribnik (szerk.) The Oxford Guide to the Uralic Languages.(Oxford Guides to the World's Languages) Oxford*, pages 674–708.

Bruno Guillaume, Kim Gerdes, Kirian Guiller, Sylvain Kahane, and Yixuan Li. 2024. Joint annotation of morphology and syntax in dependency treebanks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9568–9577.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021a. Annotation guidelines of ud and sud treebanks for spoken corpora: A proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 35–47.

Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2021b. A morph-based and a word-based treebank for beja. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 48–60.

Nikolett Mus. 2023. Nenets. In *The Uralic Languages*, pages 853–896. Routledge.

Irina Nikolaeva. 2014. *A grammar of Tundra Nenets*. Walter de Gruyter GmbH & Co KG.

Irina Nikolaeva and Edward Garrett. 2015. The endangered languages and cultures of siberia.