Creating a multi-layer Treebank for Tundra Nenets

Nikolett Mus
ELTE Hungarian Research
Centre for Linguistics
Budapest, Hungary
mus.nikolett@nytud.elte.hu

Bruno Guillaume
Université de Lorraine
CNRS, Inria, LORIA
Nancy, France
Bruno.Guillaume@loria.fr

Sylvain Kahane
Université Paris Nanterre
Modyco
Paris, France
sylvain@kahane.fr

Daniel Zeman
Univerzita Karlova
MFF, ÚFAL
Prague, Czechia
zeman@ufal.mff.cuni.cz

Abstract

This paper presents the development of the Tundra Nenets Universal Dependencies (UD) Treebank, the first syntactically annotated resource for the Samoyedic branch of the Uralic family. The treebank integrates spoken-language data and adopts the morphologically enhanced Surface-Syntactic UD (mSUD) framework to capture inflectional morphology and morphology-based syntactic relations. It further incorporates Information Structure annotation. The methodological workflow includes data selection, transcription conventions, sentence and lexeme segmentation, annotation of spoken-language features, lemmatization, treatment of morpheme status, part-of-speech and morphological tagging, and syntactic annotation based on the functional and distributional properties of syntactic elements. We also outline the principles guiding multi-level annotation and justify the theoretical choices underlying the integration of prosodic, morphological, and syntactic information.

1 Introduction

This paper presents the development of the (Tundra) Nenets (Samoyedic, Uralic) Universal Dependencies (UD) Treebank, including data selection and processing, levels of linguistic analysis, and the theoretical and methodological principles guiding the annotation. Given the early stage of development and the limited size of the corpus, the focus is on the foundational methodological approaches and theoretical decisions underlying the construction of the treebank.

Within the Uralic language family, the Finno-Ugric branch is already represented in UD by several treebanks (e.g., Finnish, Estonian, Hungarian, Komi, Udmurt), whereas the Samoyedic branch has remained absent from the data set. The inclusion of Tundra Nenets, a major Samoyedic language spoken in northwestern Siberia, therefore fills a significant gap and contributes to a more balanced coverage within the Uralic family.

Several digital corpora of Nenets exist, for instance in the Endangered Languages Archive (ELAR) and the INEL Nenets corpus (Budzisch and Wagner-Nagy, 2024), yet these resources have largely remained at the level of morphological annotation and provide limited support for syntactic analysis. Syntactic structure, in particular, remains underexplored in Tundra Nenets as well as in other Siberian Uralic and Siberian Arctic languages. The Tundra Nenets UD treebank addresses this gap by providing a systematically annotated syntactic resource, thereby enabling detailed investigations of (morpho)syntactic patterns. In addition, the treebank functions as a methodological case study for adapting the UD framework to a morphologically rich Uralic language. Its development is expected to inform both the creation of comparable resources for other Samoyedic and Siberian languages and broader discussions concerning the representation of typologically complex languages within the UD framework.

The project introduces several innovations within the UD framework (de Marneffe et al., 2021). First, since the treebank is based on spoken language data, it was necessary to determine (i) the level of transcription detail, specifically, which spoken-language-specific features should be included in the syntactic analysis, and (ii) how these phenomena should be represented, that is, the corresponding annotation principles and technical solutions. Second, the morphologically enhanced version of the Surface-Syntactic Universal Dependencies framework (Gerdes et al., 2018, 2019), the mSUD model (Guillaume et al., 2024), was adopted as the basis for annotation. This framework accommodates the rich morphological structure of Tundra Nenets and allows for an explicit representation of morphology-based syntactic relations, while remaining fully compatible with the

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 77–86
December 10-12, 2025 ©2025 Association for Computational Linguistics

UD standard. Third, the Nenets treebank includes (partial) information-structural annotation as part of a new initiative within the frame of the UniDive COST Action (CA21167), which aims to extend UD with additional layers capturing the discourse-pragmatic functions of clausal constituents.

2 The Tundra Nenets language and data

2.1 The language

Nenets is classified as a member of the Samoyedic branch of the Uralic language family. Prior to the twentieth century, linguistic descriptions generally treated Tundra and Forest Nenets as the primary dialectal varieties of the Nenets language. However, these varieties differ substantially in grammar and lexicon, and are not mutually intelligible, which justifies treating them as separate languages (Hajdú, 1968; Salminen, 1998; Burkova, 2022; Mus, 2023a). Since the current treebank includes only Tundra Nenets data, this paper focuses on that variety; with Forest Nenets materials planned for inclusion in future expansions of the corpus.

The Tundra Nenets language is spoken in the northernmost regions of the Russian Federation, primarily in the autonomous Okrugs of Nenets and Yamalo-Nenets and the Taymyrsky Dolgano-Nenetsky district. It covers an extensive Arctic area, extending from northeastern Europe to northwestern Siberia (maps illustrating these territories can be found online[1]).

The language has c. 20,000 speakers, divided into Western, Central, and Eastern dialect groups, each with local subdialects (Hajdú, 1968; Tereshchenko, 1966; Salminen, 1999; Nikolaeva, 2014; Burkova, 2022; Mus, 2023a).

It is an indigenous Arctic language and is classified as *threatened* (EGIDS 6b) (Ethnologue, 2009). Although still used in everyday oral communication across generations, speaker numbers are declining. Widespread bilingualism with Russian has led to notable lexical and structural influence.

Traditionally an oral language, Tundra Nenets achieved literacy only in the late 1920s, when a Cyrillic-based orthography was introduced (Toulouze, 1999). The writing system remains non-standardized, and several Latin-based transliteration systems are employed in scholarly contexts.

Tundra Nenets is a morphologically rich, agglutinative language. Its grammatical relations are expressed mainly by suffixes attached sequentially to the stems. Despite its agglutinative character, stem and affix alternations introduce fusional features.

The language exhibits nominative–accusative alignment (Nikolaeva, 2014): subjects are marked with the nominative case, while direct objects are typically marked with the accusative case, with limited syncretic exceptions (Hajdú, 1968; Nikolaeva, 2014). Finite verbs agree with their subjects in person and number, and may also mark agreement with objects in number when these are topical (Nikolaeva, 2014). Predicate nouns, adjectives, and certain adverbs also show agreement with their subject in person and number, and can also take the suffix of the past tense without inserting an overt copula in the predicate phrase (Nikolaeva, 2014; Hegedűs et al., 2021).

Syntactically, the language is head-final and predominantly (S)OV (Tereshchenko, 1973; Nikolaeva, 2014; Burkova, 2022; Mus, 2023a), with complements preceding their heads. *Right-dislocated* elements or *afterthoughts* occasionally occur, separated by a prosodic break and distinct intonation (Mus and Surányi, 2021, 2025). Coordination generally lacks conjunctions, while subordination is expressed through non-finite verb forms that precede the main predicate. In certain subordinate clause types, the embedded subject can trigger agreement on the non-finite verb through possessive morphology (Nikolaeva, 2014; Mus, 2023b).

2.2 The data

Written and spoken materials of Tundra Nenets are accessible in several archives and collections, including the Endangered Languages Archive (ELAR)[2], the Online Documentation of Siberian Languages[3] (Nikolaeva and Garrett, 2014), and the INEL Nenets corpus[4] (Budzisch and Wagner-Nagy, 2024). Folklore, collected during fieldwork in the region, is the most commonly represented genre in these resources. The transcription conventions, transliteration schemes, and annotation frameworks employed across these collections vary considerably and are sometimes inconsistent.

An online newspaper from the Nenets Autonomous Okrug regularly publishes articles in

[1] https://nenetsresearch.github.io/thea/tools.html

[2] https://www.elararchive.org/

[3] https://siberianlanguages.surrey.ac.uk/

[4] https://inel.corpora.uni-hamburg.de/NenetsCorpus/search

Tundra Nenets alongside Russian, providing a contemporary source of written materials in the language.[5] Additionally, a digitized text collection of approximately 500,000 tokens has been compiled and normalized, representing the written variety of the language (Mus and Metzger, 2021).

Complementing these written sources, new spoken data were collected during consultations in Moscow in 2017 with a Tundra Nenets speaker from the Yamalo-Nenets Autonomous Okrug. Rather than focusing once again on folklore, the fieldwork employed methods from Language Documentation and experimental syntax to elicit semi-controlled, naturalistic language production through interactive, goal-oriented tasks. These included a modified version of the HCRC Map Task[6], the so-called Pear Story narrative task (Chafe, 1980), and a storytelling video stimulus about reindeer herding.[7] A third elicitation type made use of picture-based story sequences, in which the speaker was asked to narrate a story depicted in a series of cartoon-style illustrations. In addition, a questionnaire was designed to prompt conversation on neutral, everyday topics, like cooking, free-time activities, public transport, and comparisons of two cities, ensuring that no sensitive personal information was collected. Finally, a set of scripted dialogues was read aloud providing controlled data for analysing prosodic phrasing and syntactic structures under comparable conditions.

The narratives were recorded in audio format (.wav), and the native speaker participant transcribed a subset of these materials orthographically using an extended Cyrillic alphabet.

Table 1 provides a summary of the available data and their current processing status from the UD perspective. Tasks and datasets that have already been processed and incorporated into the Tundra Nenets UD treebank are highlighted in green[8], while the remaining materials will be processed and added in subsequent releases of the treebank.

Type of task	Length (sec)	Nr. of sentences
HCRC Map Task 1–4	340	93
video-based storytelling: Pear Story	355	78
video-based storytelling: Arctic reindeer	235	n.d.
picture sequence narration 1–4	576	n.d.
thematic topic guided monologue 1–4	1,403	n.d.
scripted dialogue reading 1–3	232	n.d.

Table 1: Tundra Nenets spoken datasets and their current UD processing status

3 Procession of the data

In the following sections, we outline the data processing workflow, describe the methodology used to establish it, and discuss language- and data-specific annotation decisions.

3.1 Transcription and annotation of spoken language phenomena

As noted above, several texts have already been transcribed by the native speaker participant. These recordings were selected as the starting point for annotation, as sentence segmentation had already been performed by the speaker. This made the data particularly suitable for addressing one of the central theoretical challenges in spoken-language analysis: defining what constitutes a syntactic unit. The analysis of these materials provides the empirical foundation for subsequent data processing and for our working principle, which holds that intonational and semantic criteria should be jointly considered when determining sentence boundaries in spoken data.

While intonation units provide important cues for segmentation, they do not always coincide with syntactically or semantically complete utterances. Accordingly, a prosodic boundary was treated as a sentence boundary only when the preceding unit expressed a complete meaning. When an intonational unit was semantically incomplete, the subsequent material was incorporated into the same sentence. This approach ensures that sentence segmentation reflects both the prosodic organization of speech and the syntactic and semantic coherence required for UD annotation. The audio files and their transcriptions were manually annotated and time-aligned at the sentence level in Praat (Boersma and Weenink, 2025).

In addition to sentence-level prosodic alignment, individual lexemes were time-aligned with their corresponding segments in the recordings. This was done to facilitate morphological and syntactic interpretation and to support future research

[5] `https://nvinder.ru/`

[6] `https://groups.inf.ed.ac.uk/maptask/maptasknxt.html`

[7] The Pear Story task was included purely as an exploratory experiment. Given its culturally foreign context, we anticipated that the task would be challenging and highly open to interpretation, yet the language consultant completed it with remarkable fluency and engagement.

[8] To be included in release 2.17 (November 2025).

on the syntax–prosody interface. This step was also undertaken manually.

At this stage, several decisions were required concerning the treatment of spoken-language phenomena and the desired level of analytical detail. The transcriptions prepared by the native speaker follow a normalized Cyrillic orthography that reflects standardized dialectal forms rather than surface phonetic realizations. Consequently, phonetic transcriptions that capture the morphophonological peculiarities of the language were not produced at this stage. For instance, external sandhi processes – phonological alternations operating across word boundaries – observed in the language were not annotated.[9] The corresponding audio recordings, however, will be made publicly available for reference.

Instead, only those spoken-language phenomena were annotated that directly affect word-level analysis, particularly the identification of word boundaries. This decision reflects both the current lack of established UD guidelines for spoken data and the absence of detailed prosodic or phonetic descriptions of Tundra Nenets.

Rather than adopting an external prosodic annotation framework, an inductive approach was taken: recurrent lexeme-level spoken phenomena were identified directly from the recordings and transcriptions. For each such phenomenon, a dedicated annotation tag was created and, where applicable, linked to a syntactic relation within the UD framework. The conventions were inspired by existing spoken UD treebanks and preliminary annotation guidelines (Kahane et al., 2021; Dobrovoljc, 2025), but in several cases, the labeling strategy was adapted to accommodate the specific structural and typological features of Tundra Nenets. Once defined, the conventions were applied consistently across the corpus. This approach ensures that, despite the early stage of UD-based spoken-language analysis, the current representation is internally coherent and flexible enough to accommodate future standardization efforts.

Since the available recordings consist primarily of narrative monologues, certain interactional features typical of spontaneous dialogue – such as overlapping speech – are not attested.

To achieve a detailed lexical representation, two groups of spoken-language items were annotated: non-lexical and lexical items. Non-lexical items were included primarily to ensure the precise identification of word boundaries. As they do not constitute syntactic units, they were uniformly assigned the `discourse` dependency relation. This category includes:

- Noises <n> (both speaker-generated, e.g., cough, laugh, sigh, and environmental, e.g., background chatter, traffic, microphone bumps);
- Pauses <p> occurring within smaller syntactic units or between unrelated constituents;
- Audible disfluencies <d>, such as hesitation markers (“uh”, “erm”).

Lexical interruptions, by contrast, directly affect syntactic interpretation. These include:

- Unfinished lexemes <un>, which may leave grammatical relations incomplete (e.g., missing a required case marker);
- False starts <f> and repetitions (exact or partial) <er> and <pr>, which may alter expected word order;
- Incorrect word selections <iw>, where a semantically or morphologically related but unintended form is produced.

Finally, pauses coinciding with syntactic boundaries were annotated analogously to punctuation in written texts (POS `PUNCT` and relation `punct`), as they play a key role in delimiting sentence boundaries, while pauses corresponding to hesitations where analyzed analogously to disfluencies (POS `INTJ` and relation `discourse`).

The full inventory of annotated spoken-language features, together with their corresponding tags and syntactic encodings, is summarized in Table 2.

The treebank is encoded in the CoNLL-U standard format, with lexeme-level time-alignment information, indicating the onset and offset of each annotated item, stored in the MISC column.

3.2 Lemmatization, POS tagging, and morphological analysis

Building on the transcription, segmentation and time-aligned annotation of the spoken data, the subsequent stage of corpus development involved

[9] In connected speech, for example, the phrase *тына’ хадамбива / tinaʔ xadamb́iwa* ‘we killed the reindeer’ reindeer.acc.1pl kill-1pl) surfaces as *тына_кадамбива / tina_kadamb́iwa*.

Category	**Tag**	**POS**	**DEPREL**
Noise	<n>	INTJ	`discourse`
Pauses (hesitation)	<p>	INTJ	`discourse`
Audible disfluencies	<d>	INTJ	`discourse`
Unfinished lexemes	<un>	intended lexeme	`reparandum`
False Starts	<f>	intended lexeme	`reparandum`
Repetitions	<er> <pr>	intended lexeme	`reparandum`
Incorrect word	<iw>	intended lexeme	treated as if correct
Pauses (boundary)	<p>	PUNCT	`punct`

Table 2: Annotated spoken-language phenomena in the Tundra Nenets UD Treebank

the lemmatization, part-of-speech tagging, and morphological analysis of the data. Since no automated tools currently exist for Tundra Nenets, these steps were performed manually.

The process of lemmatization was guided by several theoretical decisions. First, in the absence of a unified written standard, dialectal variation was preserved in the forms of the lemma.[10] Second, during segmentation, only inflectional morphemes were detached from the stems, while the derivational morphology was left intact. Inflectional suffixes were retained in their attested surface forms and do not receive normalized lemmas. Third, linking vowels appearing at the boundary between stems and suffixes were treated as integral parts of the stem.

Part-of-speech tagging and morphological analysis were likewise carried out manually. Morphological features were segmented from the stems and glossed (cf. feature `Gloss`). In the segmentation and annotation process, only inflectional morphology was included, as inflectional markers directly contribute to syntactic relations, whereas derivational morphology was treated as part of the lexical stem. This treatment of morphology ensures consistency with the syntactic representation model adopted in the subsequent section, which integrates morphological and syntactic dependencies. The distinction between inflectional and derivational morphology was determined on the basis of descriptive and grammatical traditions established in Hajdú (1968); Nikolaeva (2014); Burkova (2022); Mus (2023a,b).

A distinction from these sources concerns the analysis of the verbal paradigm, specifically the treatment of the verbal linking suffix *-ӈа* (*-ŋa-*) that is added to certain verbal stems before agreement suffixes. Since the status of this suffix is not clear in the literature, this element was segmented from the verb stem and assigned the `AUX` POS, reflecting its auxiliary-like syntactic behavior within the clause.

In the nominal paradigm, the so-called predestinative suffix – *-да* (*-da*) – was also segmented, though it was commonly regarded a derivational element in the descriptive tradition. We propose that this morpheme may in fact participate in a syntactic relation with the noun it modifies, functioning similarly to a determiner. Accordingly, it was segmented and assigned the `det` dependency relation.

The annotation of morphosyntactic features and category labels follows the conventions of the aforementioned descriptive sources. The POS and morphological tagset was adapted from the Leipzig Glossing Rules and Abbreviations framework, with necessary modifications introduced to accommodate the specific structural and typological properties of Tundra Nenets.

To ensure consistency and uniformity across the corpus, all analyzed word forms were compiled into a reference TSV file containing the surface form, lemma, POS tag, morphological information, and translation (see Table 3). This file serves as a master inventory of annotated forms. New raw data are automatically compared against this reference using a Python script: whenever a match is found, the corresponding lemma, POS tag, and gloss information are automatically inserted into the new annotation. This procedure not only preserves consistency between previously annotated and newly added data, but also considerably accelerates the annotation process while retaining manual verification for forms not yet included in the reference file.

Form	**Lemma**	**POS**	**Gloss**
маря	мар”	NOUN	fence
-д’	_	ADP	-poss.gen.2sg

Table 3: Example excerpt from the reference TSV

3.3 mSUD annotation

To account for the complex morphological structure of Tundra Nenets and its role in expressing syntactic relations, the morphologically enhanced Surface-Syntactic Universal Dependencies (mSUD) framework (Guillaume et al., 2024) was adopted as the foundation for annotation.

[10] For example, the numeral ‘three’ occurs as *няр / ńar* in the Western dialect and as *няхар / ńaxar* in the Central and Eastern dialects.

The mSUD framework builds on the principles of the Surface-Syntactic UD model but extends it to explicitly represent morphology-based syntactic relations. It prioritizes functional heads within phrases, i.e. those constituents that determine the syntactic and distributional properties of the entire phrase, while defining dependency relations on functional and distributional grounds (Gerdes et al., 2019).

As noted, within this framework, both independent words and inflectional morphemes are systematically linked to their corresponding syntactic relations. Derivational morphology, by contrast, is not analyzed as directly contributing to syntactic dependencies. The main annotation choices adopted for Tundra Nenets suffixes are summarized in Table 4, which illustrates how distinct types of inflectional morphemes are represented and how their syntactic dependents are encoded within the mSUD relation set.

Inflection	**POS**	**mSUD DEPREL**
Number	DET	−det→ •
Case	ADP	• −comp:obj→
Possessive suffix	DET	−det:poss→ •
Predestinative suffix	DET	−det→ •
Tense suffix	AUX	• −comp:aux→
Mood suffix	AUX	• −comp:aux→
Subject agreement suffix	PRON	−subj→ •
Double agreement suffix	PRON	−subj:obj→ •
Non-finite verb suffix	AUX	• −comp:aux→

Table 4: mSUD annotation for Tundra Nenets

In the table above, the `subj:obj` relation in the verbal paradigm may require further explanation. As will be discussed in greater detail below, in Tundra Nenets transitive verbs can agree not only with their subject but also simultaneously with both their subject and object when the object is topical. Such agreement markers are typically unanalyzable portmanteau morphemes that cannot be segmented into separate units. Consequently, they were treated as a single unit and assigned the `subj:obj` dependency relation.

Because mSUD provides a more fine-grained representation of morphological and syntactic structure than standard UD, it offered a logical starting point for the development of the Tundra Nenets treebank. The annotation process therefore begins in mSUD and is subsequently converted to UD format. This direction of conversion is unidirectional: while mSUD can be reliably reduced to UD through structural simplification, the reverse conversion – from UD to mSUD – would require morphological information not encoded in UD and thus cannot be reconstructed automatically.

Annotation was carried out in a semi-automatic way using ArboratorGrew[11] (Guibon et al., 2020), which allows the creation and application of reusable rules to automate certain aspects of dependency annotation. While our corpus consists of approximately 200 sentences, the syntactic rules we employ are not probabilistic generalizations derived solely from this sample, but stable and well-established structural properties of the language. These rules are categorical (e.g., the agreement morphology on verbs that our annotation framework analyzes as subject marking) and are not subject to variation across larger datasets. Therefore, although the current rule set may be incomplete in the sense that additional rules could be added when annotating a larger corpus, the rules already formulated remain fully applicable and reliable regardless of corpus size. In other words, expanding the dataset would increase coverage but would not invalidate or contradict any existing rules, since they reflect structural facts of the language rather than artifacts of a small sample. For example, the following Grew rule was developed to attach modifying adjectives to their governing nouns and to add the dependency relation `mod` between them:

```
rule r1 {
  pattern { X[upos=ADJ]; Y[upos=NOUN];
    X < Y } without { * -> X }
  commands { add_edge Y -[mod]-> X }
}
```

This semi-automatic workflow ensures consistency across the corpus while allowing manual intervention for complex or ambiguous constructions.

3.4 Production of the UD treebank

Conversion from mSUD to UD is performed in two stages: first, mSUD is converted to SUD, and then SUD is converted to UD. Both conversions are encoded as a set of Grew (Guillaume, 2021) rules that are applied iteratively to make the necessary annotation changes. Figure 2 illustrates the process on one sentence of the treebank.

The first conversion (from mSUD to SUD) involves merging inflectional suffixes with the root word to which they are attached. Several rules are

[11] https://arborator.grew.fr/

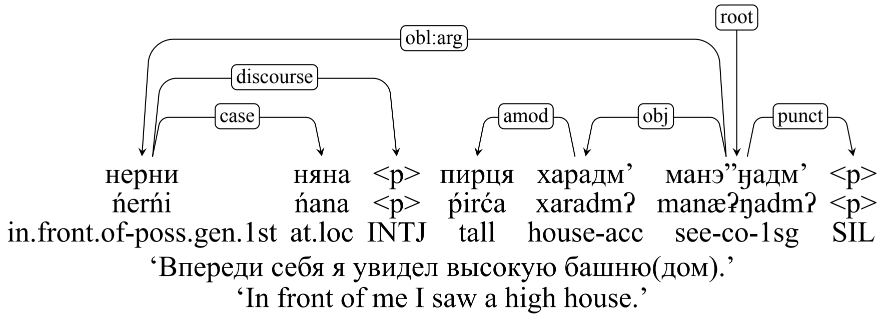
root
obl:arg
discourse
case
amod
obj
punct
нерни ńerńi in.front.of-poss.gen.1st
няна ńana at.loc
<p> <p> INTJ
пирця ṕirća tall
харадм’ xaradmʔ house-acc
манэ”ӈадм’ manæʡŋadmʔ see-co-1sg
<p> <p> SIL
‘Впереди себя я увидел высокую башню(дом).’
‘In front of me I saw a high house.’

Figure 1: Example tree from the Nenets UD treebank with Latin transliteration (2^{nd} line), English glosses (3^{rd} line) and fluent translation to Russian and English.

designed to ensure consistent glosses, lemmas and sound alignment after merging. After this first step, the word tokenization is as expected by the UD framework.

The second step relies on the general SUD to UD conversion that is described in (Gerdes et al., 2021). The syntactic structure is modified on the one hand to take into account the different choices in UD and SUD for selecting the head of a phrase in the dependency structure.[12] On the other hand, rules are used to map the SUD dependency relation tagset to the equivalent UD tagset.

The UD-annotated data produced is then validated using the process provided by the UD infrastructure.[13] These validation steps helped to identify inconsistencies in the original mSUD annotation and adapt the conversion rules to annotation choices specific to the Nenets treebank.

3.5 Information-structural roles annotation

A new initiative within the UniDive COST Action (CA21167) seeks to extend the UD framework by incorporating a layer for Information Structure (IS) annotation, drawing inspiration from the Prague Dependency Treebank 2.0.[14] In this approach, IS is treated as a functional phenomenon grounded in meaning, reflecting how speakers organize and interpret content within discourse rather than how it is formally encoded, therefore, we aim to tag IS roles in the treebank to support further formal and functional typological research.

The explicitness of IS annotation will be ensured through a detailed guideline currently under development. This guideline provides clear definitions, diagnostics, and instructions for annotators, allowing IS categories to be assigned systematically and reproducibly rather than impressionistically. Although the framework is still a work in progress and not the focus of the present paper, it reflects established best practices demonstrating that semantic and discourse-level annotation can be made reliable through well-formulated operational criteria.

In Tundra Nenets, certain IS roles are partially encoded morphologically: transitive verbs can carry suffixes marking both the person and number of the subject as well as the number of the object. Object agreement, in particular, indicates the topicality of third-person objects (Dalrymple and Nikolaeva, 2011; Nikolaeva, 2014), compare (1), where the verb agrees only with the subject, with (2), where the object is topical and the verb cross-refers its number (in addition to subject agreement).

(1) a. What did Pavel do?
Whom did Pavel see?
b. Павел Ирина-м’ манэ”ӈа-сь.
Pavel Irina-acc see.3sg-pst
‘Pavel saw Irina.’

(2) a. What did Pavel do to Irina?
b. Павел Ирина-м’ манэ”ӈа-да-сь.
Pavel Irina-acc see.3sg-sg.o-pst
‘Pavel saw Irina.’
or ‘As for Irina, Pavel saw her.’

This project adapts these insights to systematically annotate IS roles at the morpho-syntactic level, providing both a practical framework for the treebank and a model for cross-linguistic comparison.

Building on this foundation, we initiated Information Structure (IS) annotation in the Tundra Nenets UD treebank using a simple, broadly semantic scheme that captures the most fundamen-

[12] For example, the ADP is the head of the prepositional phrase it introduces in SUD, whereas in UD, this ADP depends on the main noun.

[13] https://universaldependencies.org/contributing/validation.html

[14] https://ufal.mff.cuni.cz/prague-dependency-treebank

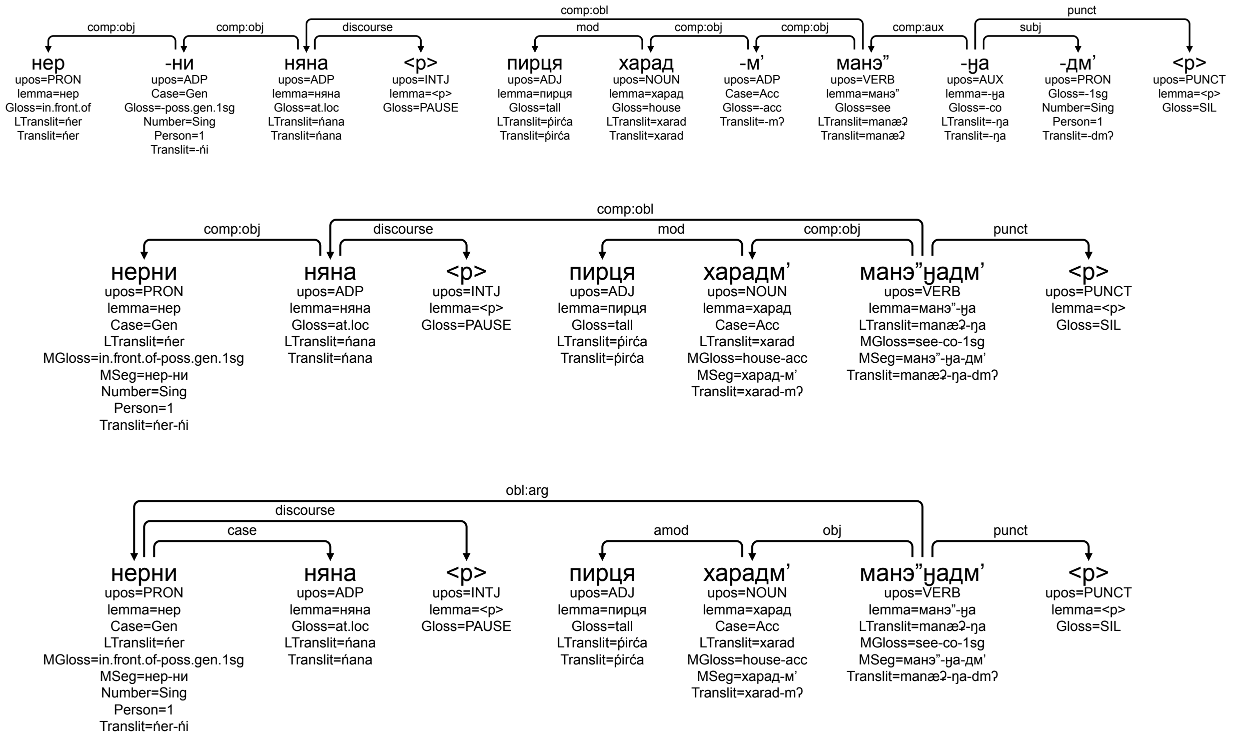

Figure 2: mSUD, SUD and UD annotations of the sentence from Figure 1.

tal distinctions observable in the corpus. We assume that IS distinctions are a universal aspect of language: all languages can differentiate contextual uses of utterances. Since these distinctions are not always overtly encoded, IS is treated primarily as a semantic phenomenon, reflecting how speakers structure and interpret information in discourse rather than as a property directly observable in form.

As a starting point, the annotation will eventually focus on topical third-person objects that trigger agreement, which are intended to be marked in the MISC column.[15] However, this annotation is not included in the current release and will be added in a future version of the treebank, once the practical framework for coding and placement in the tree is finalized.

3.6 Transliteration and translation

In addition to the Cyrillic transcription, which makes the data comparable with written Tundra Nenets resources, all annotated texts are also accompanied with a Latin-based transliteration in order to make them accessible to researchers who are not familiar with the Cyrillic script. The transliteration is generated automatically using the *translit* Perl toolkit.[16] The scheme mostly follows a 1-1 mapping between Cyrillic and Latin letters (including some characters that are not used in English, such as ŋ, æ, or the two glottal stops, ʔ and ʡ). Palatalized consonants are an exception: In the Cyrillic writing system, palatalization is often encoded in the following vowel, while in transliteration, we indicate it with an accute accent over the consonant. For example, *няна* → *ńana*.

Besides word-level English glosses, sentence-level manual translations into both English and Russian are also included. See the example tree in Figure 1.

4 Conclusions and prospects for automation

This study has presented the theoretical and methodological basis of the (Tundra) Nenets Universal Dependencies (UD) Treebank. By adapting the UD and morphologically enhanced Surface-

[15] Although one reviewer suggests that our annotation of topical objects relies only on formal features, this is not the case. In this language, object agreement is a grammatically encoded and semantically motivated marker of topicality. We use it because it directly expresses an IS value, not as a formal shortcut; the morphology itself reflects the discourse status of the argument.

[16] https://github.com/dan-zeman/translit

Syntactic UD (mSUD) frameworks to a morphologically complex Uralic language, the project established reproducible procedures for the syntactic annotation of spoken data. A particular emphasis was placed on spoken data annotation, the treatment of morphology-based syntactic relations, and the annotation of Information Structure roles. The objective of this focus was to develop a consistent and extensible annotation model.

At this stage, the Tundra Nenets treebank remains a manually annotated, small-scale resource. Subsequent endeavors will concentrate on the development of semi-automatic and fully automatic tools for lemmatization, part-of-speech tagging, and morphological analysis to facilitate corpus expansion while maintaining internal consistency and analytical precision.

Beyond its immediate scope, the project offers broader methodological insights for representing spoken and morphologically rich languages within the UD framework. The procedures and conventions developed for Tundra Nenets can be extended to other Samoyedic and Siberian Uralic languages. This contributes to a more balanced typological coverage in UD and advances the treatment of underrepresented language types in computational annotation.

Acknowledgements

This work has been supported by COST Action (CA21167) through two STSM grants: to the Faculty of Mathematics and Physics, Charles University (Beyond Syntax: Integrating Information Structure into UD Treebanks – Tundra Nenets) and to the Modyco lab, Paris Nanterre University (The syntax–prosody interface and harmonization of spoken data's syntactic annotation – Tundra Nenets). The development of the Tundra Nenets treebank was also supported by the projects Autogramm: Induction of Descriptive Grammar from Annotated Corpora (ANR-21-CE38-0017) and ThEA: Theoretical and Experimental Approaches to Dialectal Variation and Contact-Induced Change (NKFIH FK 129235), which contributed to data collection and treebank creation. Daniel Zeman's work has been supported by *LINDAT/CLARIAH-CZ*, project No. LM2023062 of the Ministry of Education, Youth, and Sports of the Czech Republic.

References

Paul Boersma and David Weenink. 2025. Praat: Doing phonetics by computer [computer program]. `https://praat.org`. Version 6.4.45, retrieved 12 October 2025.

Josefina Budzisch and Beáta Wagner-Nagy. 2024. INEL Nenets corpus. version 1.0.

Svetlana Burkova. 2022. Nenets. In Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik, editors, *The Oxford guide to the Uralic languages*, pages 674–708. Oxford University Press, Oxford.

Wallace L. Chafe, editor. 1980. *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. Advances in Discourse Processes, vol. III. Ablex, Norwood, NJ, USA.

Mary Dalrymple and Irina Nikolaeva. 2011. *Objects and Information Structure*. Cambridge Studies in Linguistics. Cambridge University Press.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Kaja Dobrovoljc. 2025. Counting trees: A treebank-driven exploration of syntactic variation in speech and writing across languages. *arXiv preprint arXiv:2505.22774*.

Ethnologue. 2009. Ethnologue: Languages of the world.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium. Association for Computational Linguistics.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2019. Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic relations and deep syntactic features. In *TLT 2019 - 18th International Workshop on Treebanks and Linguistic Theories*, pages 126–132, Paris, France. Association for Computational Linguistics.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2021. Starting a new treebank? go SUD! In *Proceedings of the Sixth International Conference on Dependency Linguistics (Depling, SyntaxFest 2021)*, pages 35–46, Sofia, Bulgaria. Association for Computational Linguistics.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *LREC 2020-12th Language Resources and Evaluation Conference*.

Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175, Online. Association for Computational Linguistics.

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, Torino, Italia. ELRA and ICCL.

Péter Hajdú. 1968. *The Samoyed peoples and languages*, volume 14 of *Indiana University Publications, Uralic and Altaic Series*. Indiana University, Bloomington. 2nd edition.

Veronika Hegedűs, Nikolett Mus, and Balázs Surányi. 2021. Tense, agreement and copula drop in Tundra Nenets copular clauses.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021. Annotation guidelines of UD and SUD treebanks for spoken corpora. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages pp–35. Association for Computational Linguistics.

Nikolett Mus. 2023a. Nenets. In Daniel Abondolo and Riitta-Liisa Valijärvi, editors, *The Uralic Languages: Second Edition*, pages 853–896. Routledge, London.

Nikolett Mus. 2023b. Tundra Nenets. In Anja Behnke and Beáta Wagner-Nagy, editors, *Clause Linkage in the Languages of the Ob-Yenisei Area. Asyndetic Constructions.*, pages 133–174. Brill.

Nikolett Mus and Réka Metzger. 2021. Toward a corpus of tundra nenets: stages and challenges in building a corpus. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 2, pages 4–9.

Nikolett Mus and Balázs Surányi. 2021. Post-verbal phrases and their correlates in Tundra Nenets.

Nikolett Mus and Balázs Surányi. 2025. Postverbal phrases in Tundra Nenets: An empirical study of rigid verb finality. Under review at *Folia Linguistica*.

Irina Nikolaeva. 2014. *A Grammar of Tundra Nenets*, volume 65 of *Mouton Grammar Library*. De Gruyter Mouton, Berlin / Boston.

Irina Nikolaeva and Edward Garrett. 2014. Online documentation of siberian languages. Audio resource.

Tapani Salminen. 1998. Nenets. In Daniel Abondolo, editor, *The Uralic Languages*, pages 516–547. Routledge, London.

Tapani Salminen. 1999. Tundra Nenets. online.

Natal'ja M. Tereshchenko. 1966. Neneckij jazyk. In V. I. Lytkin and K. E. Majtinskaja, editors, *Jazyki narodov SSSR. Volume 3: Finno-ugorskie i samodijskie jazyki*, pages 376–395. Nauka, Moscow / Leningrad.

Natal'ja M. Tereshchenko. 1973. *Sintaksis samodijskix jazykov: prostoe predloženie [The syntax of Samoyedic languages: The simple clause]*. Nauka, Leningrad.

Eva Toulouze. 1999. The beginning of a written culture by the Uralic peoples of the north. *Pro Ethnologia*, 7:52–85.