INEL Tavda Mansi Corpus

User documentation

Beáta Wagner-Nagy and Katalin Sipőcz

Hamburg, 15.05.2025

1. Introduction

1.1. Objective of the corpus

The present corpus of Tavda Mansi has been created as part of the long-term research project INEL ("*Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages*")[1] in the context of the Academies' Programme[2], coordinated by the Union of the German Academies of Sciences and Humanities[3]. Its primary goal is to create digital and machine-searchable corpora of several indigenous Northern Eurasian Languages.

The INEL Tavda Mansi corpus at hand fills a gap in the documentation of the indigenous languages of Northern Eurasia and makes possible further descriptions of the language. Mansi is a relatively good described language: there are numerous descriptions (cf. Bakró-Nagy, Sipőcz / Skribnik 2022 as well as the literature listed here) and a corpus is also available[4], however, the Tavda variety is not included in the mentioned corpus.

The analysis of materials from the Tavda variety has already been conducted by Norbert Szilágyi., but he did not produce a corpus that could be searched and evaluated electronically. However, he has made his materials available under the URL: https://norbertszilagyi91.wixsite.com/tawdamansi. In the material published in the INEL corpus, the analyses differ significantly from Szilágyi's analysis. For the sake of comparison, the texts analysed by Szilágyi are appended to the corpus, and the Hungarian translations he provided have been retained, but some places have been corrected.

The corpus that is now available can be of use not only for the study of the Mansi language, but it can be a valuable tool for both language-specific and typologically oriented research.

1.2. Mansi language

1.2.1. Description

Mansi (also known as Vogul) is spoken by appr. 939 people (VPN 2010) in Western Siberia along the river Ob and its tributaries such as Lozva, Pelym, Vagilsk, Konda, Tavda, etc. Genetically, Mansi belongs to the Ob-Ugric subgroup of the Uralic language family, its closest relatives being Khanty. In spite of its extremely low number of speakers, Mansi must be regarded as a highly endengared language.

1.2.2. Language codes

ISO 639-3 code: **mns**

Glottocode**:** **mans1258** for Northern Mansi); **cent2322** for Central Mansi; **east2879** for Eastern Mansi; **sout3253** for Southern Mansi

[1] https://www.slm.uni-hamburg.de/inel/, last access: 03.11.2021.
[2] http://www.akademienunion.de/en/research/the-academies-programme/, last access: 03.11.2021.
[3] http://www.akademienunion.de/en/, last access: 03.11.2021.
[4] https://www.babel.gwi.uni-muenchen.de/index.php?abfrage=all_mansi_corpus&subnavi=corpus_pub

1.2.3. Dialectal subdivisions

Due to its wide geographical distribution, Mansi exhibits a highly developed dialectal division. It is divided into four dialectal group (also called variety). The differences between the varieties are considerable at all levels of the language (phonetics, morphology, syntax as well as lexicon). All varieties can be subdivided into dialects. Nowadays only the North dialect is spoken. The dialects and varieties are given in Table 1.

Table 1: Distribution of the Mansi dialects and dialect groups

Variety	**Sub-dialects**				
North	Sosva	Upper Lozva	Sygva	Ob	
East	Upper Konda	Middle Konda	Lower Konda	Yukonda	
West	Middle Lozva	Lower Lozva	Northern Vagilsk	Southern Vagilsk	Pelym
South or Tavda	Chandyri	Yanychkovo	Gorodok	Shaytanskaya	Kuzyayeva

Nowadays only North Mansi is spoken in some villages along the tributaries of the rivers Ob, Sosva, Syva as well the Upper Lozva. The Western dialects were spoken along the rivers Lozva, Vagilsk and Pelym. It became extinct at the beginning of the 20th century, just like the Tavda dialect, which was last spoken in only three villages, Chandyri, Yanychkovo, and Gorodok. East Mansi is also regarded extint. In the beginning of the 20th century a handful of people spoke the language along the rivers Konda and Yukonda.

1.2.4. Tavda Mansi

Tavda Mansi was spoken along the Tavda river. The people wandered into this territory, where Siberian Tatar settled, in the 16th century. The Tavda Mansi settled among the Tatars and gradually assimilated with them. From the 17th century they were exposed not only to Tatar but also to increasing Russian influence, which led to the disappearance of the language in the early to mid 20th century. According to Kannisto (1907: 3), the number of Tavda Mansi speakers at the beginning of 20th century was about 330.

1.2.5. Research on Tavda Mansi

The research activities of Bernát Munkácsi (1860–1937) and Artturi Kannisto (1874–1943) offer a valuable opportunity to gain insight into the customs and practices of the Tavda Mansi people. Munkácsi spent a total of 11 months in 1888–1889 among the Mansi. His primary objctive was to decipher and to translate the text material collected by Antal Reguly (1819–1858). The collection of new folklore materials was considered a secondary objective. He visited three Tavda villages, where he worked with three speakers (in each village with one speaker). Mukácsi was unacquainted with this dialect prior to the research, and was surprised by the language itself, which he had initially considered to be a separate language.

Artturi Kannisto worked 4,5 years (September 1901–December 1906) among the Mansi people, but only appr. 4 month spent among Tavda people. His primary goals was to collect folklore materials and compile a dictionary as well as provide grammatical descriptions. He managed to work with eight people and he may already have been familiar with the Tavda dialect through personal communications with Munkäcsi and Munkácsi's works. The list of the consultantss is given in Table 2.

Table 2: List of the Tavda Mansi consultants

Name of the cosultant	**Abbreviation**	**Village**	**Consultant of**
two old people	–	Gorodok	Kannisto, Artturi
Kostin, Arsentej Teren`tich	KAT	Chandyri	Mukácsi, Bernát
Kostin, Logan Sidorovich	KLS	Chandyri	Kannisto, Artturi
Kostina, Aksin`ya Ivanovna	KAI	Chandyri	Kannisto, Artturi
Kostina, Tat`yana Erasimovna	KTE	Chandyri	Kannisto, Artturi
Maty`kov, Andrian Xaritonovich	MAX	Yanychkovo	Kannisto, Artturi
Matukov, Filimon Xaritonich	MFX	Yanychkovo	Mukácsi, Bernát
Maty`kov, Matvej Andrianovich	MMA	Yanychkovo	Kannisto, Artturi
Maty`kova, Alyona Filippovna	MAF	Yanychkovo	Kannisto, Artturi
Simpaev, Pavel Ignatich	SPI	Kuzyayeva	Mukácsi, Bernát
Esenbaev, Anisim Fyodorovich	EAF	Shaytanskay	Kannisto, Artturi

There are only a few works on the Tavda dialect. The descriptions proffered by Munkácsi (1894) and Honti (1975) are worthy of particular note.

1.3. Archiving

The corpus comprises annotated transcripts in *EXMARaLDA*[5] transcript formats and metadata descriptions in *EXMARaLDA* Coma format (see section 2.6 for details).

The corpus is archived and published by the Research Data Repository of the Universität Hamburg[6] under open-access conditions with Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0).[7]

The corpus is available for download in two packages of different size:

- The "standard" package includes PDF files.
- The "lite" package does not include any PDF files.

Besides the downloadable packages, the corpus is accessible online through Tsakorpus,[8] an open-source search platform for linguistic corpora (see 3.4.2 for details). The current version of the corpus can be accessed at https://inel.corpora.uni-hamburg.de/TavdaMansiCorpus/search.

1.4. Citation

The corpus is to be cited as follows:

Sipőcz, Katalin & Wagner-Nagy, Beáta. 2025. INEL Tavda Mansi Corpus. Version 1.0. Publication date 2025-05-15. https://hdl.handle.net/11022/0000-0007-FE69-6. Archived at Universität Hamburg. In: *The INEL corpora of indigenous Northern Eurasian languages.* https://hdl.handle.net/11022/0000-0007-F45A-1

Note that the authorship of the corpus refers to linguistic analysis (i.e. principally morpheme-by-morpheme glosses) of included texts. Many other people contributed to the corpus. First of all, needless to say, this refers to Mansi speakers who acted as storytellers. Furthermore, this refers to those who did the recording, transcribing, translating, annotating, editing and technical processing of the data included into the corpus. Everyone's input is acknowledged throughout the corresponding sections of this document and in the metadata.[9]

1.5. Project members

Project summary information

The INEL Tavda Mansi corpus has been developed within the long-term INEL project ("Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages"), 2016–2033. For an overview of the INEL project, see Arkhipov & Däbritz (2018). The research was carried out at the Institute for Finno-Ugric/Uralic Studies (IFUU) of the Universität Hamburg (UHH) in cooperation with the Institute of Finno-Ugric Studies of the University of Szeged.

The project homepage can be visited at: https://www.slm.uni-hamburg.de/inel/.

Project leader

Prof. Dr. Beáta Wagner-Nagy

Editors

Beáta Wagner-Nagy, Alexandre Arkhipov, Maria Brykina

Main Corpus authors

Dr. Katalin Sipőcz (University of Szeged); Prof. Dr. Beáta Wagner-Nagy (University of Hamburg)

---

[5] http://exmaralda.org/en/, last access: 03.11.2021.

[6] https://www.fdr.uni-hamburg.de/communities/inel, last access: 03.11.2021

[7] https://creativecommons.org/licenses/by-nc-sa/4.0/, last access: 03.11.2021.

[8] https://github.com/timarkh/tsakorpus, last accessed: 17.11.2024.

[9] As Sipőcz and Wagner-Nagy were the main developers of the corpus in the INEL project, they have the main responsibility for remaining errors, inconsistencies and further shortcomings of the corpus. The same applies to analytical and representation solutions, except for those following the framework of the INEL project in general.

Technical developers

Elena Lazarenko

Aleksandr Riaposov

Assistants

Richter, Felix

Petschallies, Christiana

Student Assistants

Kim, Natalia

Grell, Joshua

1.6. Acknowledgements

Fuding

This corpus has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities. The project was applied for by Prof. Dr. Beáta Wagner-Nagy, Dr. Michael Rießler, Hanna Hedeland, M.A., and Timm Lehmberg, M.A.

Legacy data

The texts in this corpus originate from published works, which are appropriately cited in the relevant sections of the metadata. The list of the publications is given in section 2.2.

2. The corpus

2.1. The language(s) of the corpus

2.1.1. Content

The corpus contains the Tavda Mansi materials collected by Bernát Munkácsi and Artturi Kannisto.

There is always only one main transcription tier (per speaker), using the common INEL transcription style (see 3.2.2).

2.1.2. Annotations

The main annotation language in the corpus is English.

Translations of the original text are provided in English, German, Hungarian and Russian (see tiers **fe**, **fg**, **fh**, **fr**). The original translation of the texts collected by Munkácsi is in Hungarian (**lth**), while the translation of Kannisto's texts is in German (**ltg**).

Morpheme glosses in English and Russian are provided for lexical items; labels for grammatical morphemes are identical in the respective tiers and are based on abbreviations of English terms, largely following Leipzig Glossing Rules (see tiers **ge, gr**).

2.1.3. Metadata

The language of metadata is English; Russian spellings of the personal names and place names are also provided in communications and speaker metadata. On transliteration of names, see 2.5.7.

2.2. Sources

The corpus contains only written data. The material of the corpus stems from the following sources:

Kannisto, Artturi and Matti Liimola 1951: *Wogulische Volksdichtung* gesammelt und übersetzt von Artturi Kannisto, bearbeitet und herausgegeben von Matti Liimola Volume I. *Texte mythischen Inhalts*. [Mémoires de la Société Finno-Ougrienne 101]. Helsinki: Suomalais-Ugrilainen Seura.

Kannisto, Artturi and Matti Liimola 1955: *Wogulische Volksdichtung* gesammelt und übersetzt von Artturi Kannisto, bearbeitet und herausgegeben von Matti Liimola Volume II*. Kriegs und Heldensagen*. [Mémoires de la Société Finno-Ougrienne 109]. Helsinki: Suomalais-Ugrilainen Seura.
Kannisto, Artturi and Matti Liimola 1956: *Wogulische Volksdichtung* gesammelt und übersetzt von Artturi Kannisto, bearbeitet und herausgegeben von Matti Liimola Volume III. *Märchen*. [Mémoires de la Société Finno-Ougrienne 111]. Helsinki: Suomalais-Ugrilainen Seura.
Kannisto, Artturi and Matti Liimola 1958: *Wogulische Volksdichtung* gesammelt und übersetzt von Artturi Kannisto, bearbeitet und herausgegeben von Matti Liimola Volume IV*. Bärenlieder*. [Mémoires de la Société Finno-Ougrienne 114]. Helsinki: Suomalais-Ugrilainen Seura.
Kannisto, Artturi and Matti Liimola 1963: *Wogulische Volksdichtung* gesammelt und übersetzt von Artturi Kannisto, bearbeitet und herausgegeben von Matti Liimola Volume VI. *Schicksalslieder, Klagelieder, Kinderreime, Rätsel, Verschiedenes*. [Mémoires de la Société Finno-Ougrienne 134]. Helsinki: Suomalais-Ugrilainen Seura.
Munkácsi, Bernát 1896: *Vogul népköltési gyűjtemény* IV. Életképek. Budapest: Magyar Tudományos Akadémia.

2.3. Content

The corpus contains texts of various genres, which are classified (following the INEL project conventions) as folklore, narrative (monologues that are neither folklore texts nor songs), conversation and song.

2.4. Corpus size

The corpus currently contains 29 transcripts of 11 speakers with 2,042 utterances and 11,879 tokens.

2.5. Naming conventions

2.5.1. Name of the corpus

The name of the corpus is INEL Tavda Mansi corpus.

2.5.2. Folder structure and file names

The entire corpus is contained in the folder “tavda-mansi” which has the following files and subfolders.

2.5.3. Folders with text transcripts, organized by genre:

- “flk” (folklore texts)
- “nar” (narrative texts)
- “sng” (songs)
- “conv” (conversation)

Each of these genre folders contains one further subfolder per text (“communication”), named identically to the text name (see 2.5.4). Each text folder contains several files with different extensions according to the file type (see 2.6 for details on file formats):

- annotated transcript in EXMARaLDA EXB and EXS formats (*.exb, *_s.exs)
- annotated transcript converted into ELAN[10] format (*.eaf)
- annotated transcript converted into ISO/TEI format (*_tei.xml)
- (if available) original version coming from published sources in PDF (*.pdf) format (scanned images or text depending on the available source)

Supplementary folders:

- "documentation" (contains the present document)
- "corpus-utilities" (contains annotation panel files that can be opened in EXMARaLDA Partitur Editor):
- "annotation-panel-inel.xml": annotation values (along with short descriptions) used in tiers SeR, SyF, IST, BOR, BOR-Phon, BOR-Morph, CS, ExLocPoss (in this and other currently developed INEL corpora; thus includes values not encountered in the present corpus; see 3.3.8–3.3.14)
- "gloss-panel-tavda-mansi.xml": annotation values used in the part-of-speech tier (**ps**, see 3.3.6) and glossing labels for grammatical meanings used in tiers **ge**, **gr** (see 3.3.4), along with short descriptions

[10] https://tla.mpi.nl/tools/tla-tools/elan/, last access: 07.05.2025

Individual files:

- "tavda-mansi.coma" (main metadata file; see 2.8)
- "coma_overview.html" (a browser-readable overview of the main metadata file)

2.5.4. Names of texts (communications)

The names of the texts which are used as their IDs throughout the corpus are composed of the following components: main speaker code (see 2.5.5); further speaker codes (optional); year of recording; short title; genre abbreviation. These components are joined by underscore ("_").

The short title as part of a text name is a (possibly shortened) version of the English title, spelled without spaces, hyphens or other non-letter characters, with all initial capitals.

The genre abbreviations used are listed in 2.5.6.

In what follows an example of a text name can be seen:

Communication code: EAF_19031221_TaleOfPokatikorokh_flk

Speaker code: EAF (Esenbaev, Anisim Fyodorovich)

Date of recording: 1903.12.21.

Short title: TaleOfPokatikorokh (i.e. “Tale of Pokatikorokh”)

Genre: flk (folklore)

2.5.5. Speaker codes

The speaker codes are derived from the speaker’s full names in the order “Family name — First name — Patronymic” in their INEL Latin transliteration. Most commonly, a code is thus composed of three initial capital letters, e.g. “EAF” stands for Esenbaev, Anisim Fyodorovich (Есенбаев, Анисим Фёдорович). Table 2 in section 1.2.5 contains the full list of Tavda Mansi speakers who contributed to the corpus along with their codes.

2.5.6. Abbreviations used in metadata

DCh: Däbritz, Chris Lasse

GJ: Grell, Joshua

KiN: Kim, Natalia

PCh: Petschallies, Christiana

SK: Sipőcz, Katalin

WNB: Wagner-Nagy, Beáta

2.5.7. Transliteration of Cyrillic names

In the metadata fields referring to personal names and placenames, a romanized spelling is used alongside their Cyrillic spelling according to the Russian orthography.

All personal names and most placenames in the metadata are transliterated following the GOST 7.79–2000 System B transliteration standard (GOST 2001). Elsewhere, e.g. in text titles, English glosses and free translation, English-style romanization is used.

2.6. Technical formats

2.6.1. Transcripts

The annotated transcripts are delivered in the formats of the EXMARaLDA software suite, all of them in XML. The main transcript file which can be used for browsing the transcript with the EXMARaLDA Partitur Editor is the “basic transcription” format (EXB). From the basic transcription, a supplementary “segmented transcription” (EXS) is automatically generated which is necessary to make searches across the corpus with the EXMARaLDA EXAKT corpus search tool and to provide word and sentence counts. (Note that the segmented transcription files are **not** to be opened with the Partitur Editor.) The respective file extensions are “.exb” and “.exs”. Files encoded in the ISO/TEI standard for “Transcription of Spoken Language” (file extensions is “.xml”) are intended to be used for enhanced interoperability and export.

Additionally, the annotated transcripts are converted into ELAN format (“.eaf”), which makes the downloaded corpus also browsable and searchable locally using ELAN. ELAN transcripts differ from the original EXB transcripts

in tier structure due to inherent differences between the two data models. In EXB transcripts, the main transcription tier is the tier tx (with subdivision into words), and all other tiers are dependent on tx. In ELAN transcripts, the main transcription tier is the tier ts (sentence-level), and all other tiers are dependent on ts. Please be aware that the ELAN versions of the transcripts are provided for compatibility only and are not specifically tested or curated.

2.6.2. Metadata

The corpus metadata are created in the EXMARaLDA Coma (corpus manager) and stored in the Coma XML format (file extension ".coma"). One file holds the metadata for the whole corpus.

2.6.3. Media

For some texts the corresponding pages were scanned and are provided in PDF format (file extension ".pdf").

2.6.4. Other data

No other data types are provided with the corpus.

2.7. Workflow of the source files

2.7.1. Transcripts

All texts were typed and converted to a simplified standard phonemic transcription and then imported into *SIL Fieldworks Language Explorer* (FLEx)[11] for glossing.

For all transcripts, the morphological analysis (interlinear glossing) was done in FLEx. This is where all the morpheme-level tiers were created (*mb, mp, ge, gr, mc*), as well as the part-of-speech tier (ps). The BOR tier was also pre-filled directly from the FLEx lexicon.

As soon as glossing is complete, a text is exported from FLEx as flextext XML and converted to EXMARaLDA EXB format. During this conversion, the *ref* tier is created which combines communication code and sentence numbering (see below). There are also some changes to the *tx* tier concerning punctuation and to the morpheme-level tiers concerning the representation of zero morphs (see below). After that, all further annotating and editing is done in the EXMARaLDA Partitur-Editor.

2.7.2. Metadata

The metadata of the corpus are stored in *EXMARaLDA Coma* format. It is an XML-based format with separate interlinked descriptions for communications (texts; also analogous to IMDI "sessions") and speakers. The fields contained in the descriptions are listed in the following sections. This includes for example the location and date of a communication, but also information on which part of the processing and analysis was done by whom. Metadata about speakers contains mainly biographical data, but also basic data on language proficiency.

2.8. Metadata for the corpus

2.8.1. General corpus metadata

The general metadata about the whole corpus include the corpus name ("INEL Tavda Mansi Corpus") and some basic metadata fields complying with the standards of DC (Dublin Core) and OLAC (Open Language Archive Community).

2.8.2. Communication metadata

Name: The code which is assigned to the text (see 2.5.4)

Description:

- **0a Title:** Complete title of the communication.
- **0b Title (RUS):** Complete title of the communication in Russian.
- **1 Genre:** Abbreviation of the genre of the communication (conv = conversation, flk = folklore, nar = narrative, sng = song).
- **2a Recorded by**: The full name of the person by whom the communication was recorded (either Bernát Munkácsi or Artturi Kannisto).
- **2b Date of recording:** Here the date of recording is given (year only).

[11] https://software.sil.org/fieldworks/, last access: 11.04.2025.

- **3a Dialect group:** Here the dialect (Tavda) is given, where the transcript comes from.
- **3b Dialect:** Here the subdialect (Chandyri, Kuzyayeva, Shaytanskaya) is specified, where the transcript comes from.
- **4 Transcribed by:** Full name of the person who did the transcription (eiter Bernát Munkácsi or Artturi Kannisto).
- **5a Date of transcribing:** The year of the transcribing.
- **5b Speakers:** Code(s) of the speaker(s).
- **5c Typed by:** The name of the person who typed the text.
- **5d Time-aligned:** The name of the person who aligned the text with the audio. Tis tier is not relevant for this corpus. The tier is always filled as no sound.
- **6a-g Translation(s):** Abbreviation of the person who did the translation in question (Russian, English, German, Hungarian).
- **7a-f Annotation SeR / SyF / IST / BOR_CS / ExLocPoss / geo:** Codes of the persons who annotated the respective tiers (SeR; SyF; IST; BOR, BOR-Phon, BOR-Morph, CS, ExLocPoss and geo; see 2.5.6).
- **8a Glossed by:** Abbreviation of the person who did the glossing.
- **8b Gloss checked by:** Abbreviation of the person who checked the glossing.
-

Location: The following fields specify the location where the text was recorded.

- **Country:** The country where the recording took place; this is always Russia.
- **Region:** The region/administrative unit where the recording took place. We indicate the administrative unit at the time of the recording – consequently.
- **Settlement (LatLng):** The longitude and latitude of the settlement where the recording took place.
- **Settlement:** The settlement where the recording took place. If no exact settlement is known, also the name of a river, a lake or a mountain ridge can be given.
- **Settlement (RU):** The Russian name of the settlement where the recording took place

Languages:

- **Language code:** The ISO language code of the communication (*mns* – Mansi).

Setting: In this section some information about archive sources and existing publications is given.

- **Published in:** The publication reference is provided here
- **Published in (bibtex):** The BiBTeX key of the corresponding entry in the INEL Bibliography is given here.
- **Published online:** Link to the online material published by Szilágyi is given here.

Recording: If an audio file is available, it is linked to the communication description. This is not relevant in the Tavda Mansi Corpus.

Transcriptions: The basic transcription (.exb) and the segmented transcription (.exs) are linked here to the communication description; the latter is needed for searching the corpus.

Attached file(s): If there are additional files (e.g. scans of text publications), they are linked to the communication description here. The analysis published by Szilágyi (in pdf format) is added to the corpus here.

2.8.3. Speaker metadata

Metadata about the speaker(s) involved in a communication includes both biographical information about the speaker and information about his/her sociolinguistic background. However, due to the wide variety of communications and speakers, it is not always possible to provide detailed speaker metadata. The following information is given as accurately as possible:

Description of speaker:

- **1a Family name:** Family name of the speaker (Latin script).
- **1b Family name (RU):** Family name of the speaker (Cyrillic script).
- **2a Given name:** Given name of the speaker (Latin script).
- **2b Given name (RU):** Given name of the speaker (Cyrillic script).
- **3a Patronymic:** Patronymic of the speaker (Latin script).

- **3b Patronymic (RU):** Patronymic of the speaker (Cyrillic script).
- **4 Also known as (mns name):** Here the Mansi name is given, if it is known.

Basic biographical data: Here basic biographical data of the speaker is provided.

- **1 Place of birth:** Place of birth of the speaker (Latin script).
- **2 Region:** Region where the speaker was born.
- **3 Country**: Country where the speaker was born; this is always Russia.
- **4 Date of birth:** The speaker's date of birth.
- **5 Date of death:** If the speaker already died, the speaker's date of death.
- **6 Grown up in / former residences:** Former residences of the speaker (Latin script).
- **7 Domicile**: Location where the speaker lived at the time of the recording (Latin script).

Education: Here information – if available – is given on the speaker's education and occupation/profession

- **1 Education:** Here information on basic education (i.e. school) of the speaker is given (English).
- **2 Higher education:** If the speaker has had higher education, it is mentioned here (English).
- **3 Occupation:** Here the profession and/or occupation of the speaker is mentioned (English).

Family: Here information about the ethnicity of the respective speaker and his/her family members is given.

- **1 Ethnicity:** Ethnicity of the speaker.
- **2 Ethnicity of mother:** Ethnicity of the speaker's mother.
- **3 Name of mother:** Name of the speaker's mother.
- **4 Ethnicity of father:** Ethnicity of the speaker's father.
- **5 Name of father:** Name of the speaker's father.
- **6 Ethnicity of husband/wife:** Ethnicity of the speaker's husband/wife.
- **7 Name of husband/wife:** Name of the speaker's husband/wife.
- **8 Ethnicity of grandparents:** Ethnicity of the speaker's grandparents.
- **9 Other information:** Other relevant information about the speaker

Language documentation activities: Here it is indicated how the speakers was integrated into language documentation

- **Consultant of:** Here it is mentioned with which linguist(s) the speaker worked (either Munkácsi or Kannisto).

Languages: Here we give the language codes (*mns* notes Mansi, *rus* Russian) for the languages the speaker has command of.

- **L1**
    - **1 First language:** The speaker's first language.
    - **2 Dialect group:** The dialect group of the speaker's first language (always Tavda).
    - **3 Dialect:** The dialect of the speaker's first language.
    - **4 Subdialect: The subdialect of the speaker's first language**
- **L2**
    - **Second language:** The speaker's second language.

3. Transcription and annotation

A significant number of concepts and principles related to transcription and annotation have their origins in the *Nganasan Spoken Language Corpus* (NSLC) (Brykina et al., 2018 The documentation of these principles can be found in the respective user guidelines (Wagner-Nagy et al., 2018). This assertion is particularly salient in the context of annotation principles and annotation schemes for the annotation of semantic roles (SeR), syntactic functions (SyF) and information status (IST), as will be demonstrated in the subsequent sections. For a comprehensive overview of general principles of transcription, annotation and translation, please refer to Arkhipov (2020).

3.1. Tier layout

Every annotation tier has a distinct label (see left column in the table) which is shown in the respective EXB file. Table (3) shows all occurring tiers and gives a short description of them.

Table 3: Overview of annotation tiers

Tier label	Tier name	Description	Unit	Optionality
ref	Reference	Text ID + sentence number	sentence	obligatory
ts	Text (sentence)	Main transcription	sentence	obligatory
tx	Text (word)	Main transcription segmented by word for interlinearization	word	obligatory
mb	Morpheme breaks	Morpheme breakdown of words	morph	obligatory
mp	Morphophonemes (underlying)	Underlying (lexical) forms of morphemes	morph	obligatory
ge	Gloss (English)	Morpheme glosses (with lexical glosses in English)	morph	obligatory
gr	Gloss (Russian)	Morpheme glosses (with lexical glosses in Russian)	morph	obligatory
mc	Morphological category	Morphological category/part of speech for each morpheme	morph	obligatory
ps	Part of speech	Part of speech for each word	word	obligatory
SeR	Semantic Role	Semantic (thematic) roles for major NPs	word	optional
SyF	Syntactic function	Syntactic functions for predicates and arguments	word	optional
IST	Information status	Information status for major NPs (given/new/accessible)	word	optional
BOR	Borrowing	Borrowings (source language and type)	word	optional
BOR-phon	Borrowing phonology	Phonological adaptations in borrowings	word	optional
BOR-morph	Borrowing morphology	Morphological adaptations in borrowings	word	optional
BOR-Scr	Borrowing source	Source form of the borrowed word	word	optional
CS	Code switching	Code switching and calques (source language and type)	group of words	optional
geo	geographical info	The longitude and latitude of the place name	group of numerals	optional
fe	Free translation (English)	Free translation (English)	sentence	obligatory
fg	Free translation (German)	Free translation (German)	sentence	obligatory
fr	Free translation (Russian)	Free translation (Russian)	sentence	obligatory
fh	Free translation (Hungarian)	Free translation (Hungarian)	sentence	obligatory
ltg	Literal translation (German)	Original translation in Kannisto's works,	sentence	optional
lth	Literal translation (Hungarian)	Original translation in Munkácsi's works	sentence	optional
nt	Notes	Notes from corpus developer	sentence	optional

3.2. Transcription tiers

3.2.1. Main transcription tiers (tx, ts)

The tier **ts** presents transcriptions of entire sentences. The transcription tier (**tx**) is the most important tier in the transcriptions, as it contains the main transcription segmented into words and is the basis for all further annotations. The following example shows the **ts** and **tx** tiers.

(1)

ref	KZE_19040108_Dialog_conv.009 (001.009)
ts	wɔɔrpɛŋn noŋšɔŋgʰɔɔs.
tx	wɔɔrpɛŋn noŋšɔŋgʰɔɔs.
fe[12]	It jumped up on the top of a spruce tree.

3.2.2. Transcription and orthographical conventions in the corpus

For transliteration of Cyrillic names in the metadata, see 2.5.7.

INEL transcription system

The transcription employed in the corpus is an interpretation of the original transcription. It thus differs significantly from Kannisto's transcription.

The choice of symbols largely follows the general conventions adopted in the INEL project. In particular,

- *ŋ* is used for the voiced velar nasal
- *t', d', n', l',* are used for palatalized or palatal consonants
- *č* is used for the postalveolar/palatalized affricate
- *š* is used for the postalveolar/palatalized fricative
- double characters are used for long vowels, e.g. *oo* for /oː/
- ʰ is used to mark a spirantization

The project transcription is represented in tiers **ts** (Text (Sentence)) and **tx** (Text (Word)).

3.3. Annotation tiers

3.3.1. Reference (ref)

The reference tier (**ref**) for each sentence contains the code of the communication and the number of the sentence, separated by dot. The sentences are numbered through the entire text. The sentence numbers are zero-padded up to 3 digits. In brackets, the numbering according to the FLEx scheme is given (*paragraph_number.sentence_number*).

(2)

ref	KZE_19040108_Dialog_conv.009 (001.009)		KZE_19040108_Dialog_conv.010 (001.010)		
ts	wɔɔrpɛŋn noŋšɔŋgʰɔɔs.		wɔɔrpɛŋgʰən kɔɔt ɔɔl?		
tx	wɔɔrpɛŋn	noŋšɔŋgʰɔɔs.	wɔɔrpɛŋgʰən	kɔɔt	ɔɔl?
fe	It jumped up on the top of a spruce tree.		Where is the top of a spruce?		

3.3.2. Morpheme breaks (mb)

The morpheme breaks tier (**mb**) breaks words into segmentable morphemes. Each word – according to the tier **tx** – appears in a separate cell. The morphemes are still represented with their surface structure and are separated from each other by hyphens. Zero morphs are not represented in this tier. Productive derivational suffixes are segmented, while non-productive derivational suffixes are mostly not segmented. In this case, the derived stem is then glossed as a separate lexical item.

(3)

ref	KZE_19040108_Dialog_conv.009 (001.009)	
ts	wɔɔrpɛŋn noŋšɔŋgʰɔɔs.	
tx	wɔɔrpɛŋn	noŋšɔŋgʰɔɔs.
mb	wɔɔr-pɛŋ-n	noŋ-šɔŋgʰ-ɔɔ-s
fe	It jumped up on the top of a spruce tree.	

[12] "fe" stands for 'free English translation' (see 3.3.15). It is introduced already here in order to make the examples understandable.

3.3.3. Morphophonemes (underlying) (mp)

The underlying morphemes tier (**mp**) shows the deep structure of the morphemes which were separated from each other in **mb**. Stems are, thus, represented here by their lexical entry in the FLEx lexicon. All morphemes within a word are separated by hyphens. Zero morphs are not represented in **mp**.

(4)

ref	KZE_19040108_Dialog_conv.009 (001.009)	
ts	wɔɔrpɛŋn noŋšɔŋgʰɔɔs.	
tx	wɔɔrpɛŋn	noŋšɔŋgʰɔɔs.
mb	wɔɔr-pɛŋ-n	noŋ-šɔŋgʰ-ɔɔ-s
mp	wɔr-pöŋ-nɔ	noŋ-šɔŋk-ə-s
fe	It jumped up on the top of a spruce tree.	

3.3.4. Gloss (ge and gr)

The gloss tiers (**ge** and **gr**) contain the English and Russian glossing of the morphemes in **mb** and **mp**. Stems receive their respective lexical glosses in the two languages, while affixes are glossed identically in latin script and mostly according to the Leipzig Glossing Rules[13]. For the list of abbreviations used and the list of affixes occurring in the corpus, see Appendix 1 and Section 3.3.5, respectively. Glosses for all morphemes within a word are separated with hyphens. Non-overt morphemes are given in square brackets preceded by a dot (e.g. ".[NOM]").

If a morpheme contains two or more semantic components, then they are separated by a dot, for more convenient reading that does not hold true for the combination of person and number (e.g. IMP.2SG). The order of the semantic components is:

- mood – person/number: IMP.2SG (imperative, 2nd person singular)
- non-finite form – specification of the form: PTCP.PST (past participle

Morphemes with unknown meaning are glossed with two percent signs (%%). Morphemes, which apparently are derivational suffixes, but those function cannot be further determined, are glossed with DRV.

(5)

ref	KZE_19040108_Dialog_conv.009 (001.009)	
ts	wɔɔrpɛŋn noŋšɔŋgʰɔɔs.	
tx	wɔɔrpɛŋn	noŋšɔŋgʰɔɔs.
mb	wɔɔr-pɛŋ-n	noŋ-šɔŋgʰ-ɔɔ-s
mp	wɔr-pöŋ-nɔ	noŋ-šɔŋk-ə-s
ge	forest-head-LAT	up-jump-EP-PST.[3SG.S]
gr	лес-голова-LAT	в-прыгать-EP-PST.[3SG.S]
fe	It jumped up on the top of a spruce tree.	

3.3.5. Morphological category (mc)

The morphological category (mc) tier indicates the morphological category of both lexical stems and affixes (i.e. the inflectional category or the derivational process). Table 4 and Table 5 show the tags used for lexical stems and inflectional categories; derivational processes are marked as x > y, x and y being the tags for lexical stems. The morphological category of zero morphs is once more indicated within square brackets. Categories in brackets are optional morphems.

Table 4: Tags for lexical stems

Tag	Description
adj	adjective
adv	adverb
cl	clitic
conj	conjunction
cop	cop

[13] https://www.eva.mpg.de/lingua/resources/glossing-rules.php, last access: 04.11.2021.

cvb	converb
dem	demonstrative pronoun
inter	interrogative pronoun
intj	interjection
n	noun
npr	proper noun
num	numeral
pp	postposition
pro	pronoun
prf	prefix
prv	preverb
ptcl	particle
ptcp	participle
quant	quantifier
reln	relational noun
v	verb
%%	unknown part of speech

Table 5: Tags for inflectional categories

Tag	**Comment**
Inflection of nominals	
n:(case)	case suffix at nouns (also at adjectives, numerals, participles and pronouns)
n:(ins)	epenthetic vowel at nouns (also at adjectives, numerals, participles and pronouns)
n:(num)	number suffix at nouns (also at adjectives, numerals, participles and pronouns)
n:(poss)	possessive suffix at nouns (also at adjectives, numerals, participles and pronouns)
Inflection of verbs	
v:(ins)	epenthetic vowel at verbs
v:(mood)	mood suffix at verbs
v:(nfin)	infinitive suffix at verbs
v:(pn)	person-number suffix at verbs
v:(tense)	tense suffix at verbs
v:(voice)	voice marker at verbs

Chart (6) shows an example of how morpheme classes are represented. Categories in square brackets are zero morphems.

(6)

ref	KZE_19040108_Dialog_conv.009 (001.009)	
ts	wɔɔrpɛŋn noŋšɔŋgʰɔɔs.	
tx	wɔɔrpɛŋn	noŋšɔŋgʰɔɔs.
mb	wɔɔr-pɛŋ-n	noŋ-šɔŋgʰ-ɔɔ-s
mp	wɔr-pöŋ-nɔ	noŋ-šɔŋk-ə-s
ge	forest-head-LAT	up-jump-EP-PST.[3SG.S]
gr	лес-голова-LAT	в-прыгать-EP-PST.[3SG.S]
mc	n-n-n:(case)	prv-v-v:(ins)-v:(tense).[v:(pn)]
fe	It jumped up on the top of a spruce tree.	

3.3.6. Part of speech (ps)

The part of speech tier (**ps**) contains information about the grammatical category of each word form. Hence, e.g. the outcome of derivational processes is marked here. The tags used are more or less the same as in the morphological category tier **mc**.

(7)

ref	KZE_19040108_Dialog_conv.009 (001.009)	
ts	wɔɔrpɛŋn noŋšɔŋg^{h}ɔɔs.	
tx	wɔɔrpɛŋn	noŋšɔŋg^{h}ɔɔs.
mb	wɔɔr-pɛŋ-n	noŋ-šɔŋg^{h}-ɔɔ-s
mp	wɔr-pöŋ-nɔ	noŋ-šɔŋk-ə-s
ge	forest-head-LAT	up-jump-EP-PST.[3SG.S]
gr	лес-голова-LAT	в-прыгать-EP-PST.[3SG.S]
mc	n-n-n:(case)	prv-v-v:(ins)-v:(tense).[v:(pn)]
ps	n	v
fe	It jumped up on the top of a spruce tree.	

3.3.7. Geographical coordinates (geo)

Geographical coordinates in "latitude, longitude" format are provided in the **geo** layer for some of the place names found in the corpus.

(8)

ref	MAX_19031217_VillageOfYevsekova_flk.001 (001.001)			
ts	Sɛwlənpɔwlt ajalt Jɛpsej aals.			
tx	Sɛwlənpɔwlt	ajalt	Jɛpsej	aals
mp	Sɛwlənpɔwl-ta	ajɔl-ta	Jɛpsej	ɔl-s
ge	Yevseykova-LOC	front-LOC	Yevsey.[NOM.SG]	live-PST.[3SG.S]
geo	57.964080, 65.525686			
fe	In the village of Yevseykova, there lived first and foremost [a man named] Yevsey..			

3.3.8. Syntactic function (SyF)

In the Syntactic Function tier (**SyF**) basic syntactic functions (i.e. subject, direct object, predicate) are annotated. The annotation is also based on GRAID principles (Haig & Schnell 2014), and the annotation scheme used was developed by Beáta Wagner-Nagy and Sándor Szeverényi (Wagner-Nagy et al. 2018: 24ff.) who also made it available for the project. Hence, the tags are likewise built up according to the <form.animacy:syntactic function> scheme. Subjects and direct objects are tagged at the beginning of the sentence, null subjects are tagged at the predicate of the clause. In complex predicates, the auxiliary verb is tagged. The following tags are used to annotate syntactic functions:

Table 6: Tags for annotating syntactic functions

Abbreviation	**Comment**
Subject	
pro.h:S	pronominal human subject
pro:S	pronominal non-human subject
np.h:S	nominal human subject
np:S	nominal non-human subject
0.1.h:S	zero/covert first-person human subject
0.2.h:S	zero/covert second-person human subject
0.3.h:S	zero/covert third-person human subject
0.3:S	zero/covert third-person non-human subject
Direct Object	
pro.h:O	pronominal human direct object
pro:O	pronominal non-human direct object
np.h:O	nominal human direct object
np:O	nominal non-human direct object

Predicate	
v:pred	verbal predicate
ptcp:pred	participle predicate
n:pred	nominal predicate
adj:pred	attributive/adjectival predicate
pro:pred	pronominal predicate
ptcl:pred	particle predicate

Syntactic functions are only tagged in main clauses. Dependent/subordinate clauses are tagged separately, the cells belonging to the subordinate clause are merged. The tags are as follows:

Table 7: Tags for annotating subordinate clauses

Abbreviation	**Comment**
s:comp	complement clause (*I know that he goes.*)
s:rel	relative clause (*I know the man who is going home.*)
s:temp	temporal clause (*When I came home, nobody was there.*)
s:cond	conditional clause (*If he goes home now, I am really upset.*)
s:adv	adverbial clause (*He went home laughing loudly.*)
s:purp	purpose clause (*He went home to feed his cat.*)

The following charts show some examples of tagging syntactic functions.

(8)

ref	MAX_19031217_VillageOfYevsekova_flk.001 (001.001)			
ts	Sɛwlənpɔwlt ajalt Jɛpsej aals.			
tx	Sɛwlənpɔwlt	ajalt	Jɛpsej	aals
mp	Sɛwlənpɔwl-ta	ajɔl-ta	Jɛpsej	ɔl-s
ge	Yevseykova-LOC	front-LOC	Yevsey.[NOM.SG]	live-PST.[3SG.S]
SyF			np.h:S	v:pred
fe	In the village of Yevseykova, there lived first and foremost [a man named] Yevsey..			

(9)

ref	MAX_19031217_VillageOfYevsekova_flk.003 (001.003)					
ts	üx pɔɔst min'äs Tɛüt paaxən wɛnplüŋ.					
tx	üx	pɔɔst	min'äs	Tɛüt	paaxən	wɛnplüŋ.
mp	ük	pɔɔs-ta	min'-ə-s	Täüt	pɔɔk-ə-nɔ	wänp-l-uŋ
ge	one	time-LOC	go-EP-PST.[3SG.S]	go-EP-PST.[3SG.S]	shore-EP-LAT	ishing_rod-VBLZ-INF
SyF			0.3.h:S v:pred			s:purp
fe	Once he went to Tavda to fish.					

3.3.9. Semantic roles (SeR)

The Semantic Roles tier (**SeR**) contains the annotation of semantic roles (a.k.a. thematic roles, theta-roles). The annotation is based on GRAID principles (cf. Haig & Schnell 2014), and the annotation scheme used was developed by Beáta Wagner-Nagy and Sándor Szeverényi (Wagner-Nagy et al. 2018: 21ff.) who also made it available for the project. The annotation takes into account form, animacy and semantic role of the referent, the tags are built up according to the scheme <form.animacy:semantic role>. If the referent is expressed by a complex phrase, then the semantic role is tagged at the head of the phrase. Zero referents are tagged per default at the predicate of the sentence. Semantic roles are tagged both in main and in dependent clauses. In the category "animacy" human and non-human referents are differentiated. Human referents get the abbreviation <h>, non-human referents get no marking in this category.
The following tags for the form of the referent are used:

Table 8: Abbreviations for form of the referent

Abbreviation	**Comment**
0.1.	zero/covert first-person referent
0.2.	zero/covert second-person referent
0.3.	zero/covert third-person referent
.h	human referent
adv	adverbial referent
np	nominal referent (noun phrase)
pp	postpositional phrase
pro	pronominal referent

The semantic roles which are tagged are explained in the following table:

Table 9: Semantic Roles tagged and their abbreviations

Semantic Role	**Abbreviation**	**Comment**
Agent	A	- volitional initiator of the action - the participant which is volitionally causing the action - can be both animate and inanimate - test agent vs. theme: add "on purpose" to the sentence - if it fits, then it is an agent, if not, then not
Beneficiary	B	- entity for whose benefit the action is performed
Cause	Cau	- entity (mostly non-human) that causes an event
Comitative	Com	- entity that convoys a participant of the action (a.k.a. as co-agent)
Experiencer	E	- entity that experiences the action or event - does not have a control over the action or event - verba sentiendi, i.e. verbs expressing emotion, volition, cognition, perception (i.e. verbs like: *see, love, hate, understand, hear, taste, frighten, wish, want, think, remember, feel*)
Goal	G	- location or entity in the direction of which something moves (i.e. directional location)
Instrument	Ins	- medium by which the action or event is performed
Location	L	- location or entity where an event takes or place or where something is located (i.e. stative location)
Path	Path	- entity or location along or through which the event takes place
Patient	P	- undergoer of the action - test patient vs. theme: does the referent change its quality during the action? – if yes, then patient - first argument of unaccusative verbs such as *die, fall*
Possessor	Poss	- entity which owns something - both alienable and inalienable possession - also inanimate referents (e.g. the top of the mountain)
Recipient	R	- (mostly animate) recipient of transfer of something - addressee of speech verbs
Source	So	- location or entity where a movement starts (i.e. directional location) - original owner in a transfer of something
Stimulus	St	- stimulus for physical perception, i.e. second argument of verbs like *see*, *hear*, *feel*, but NOT of verbs like *look for*, *listen*
Theme	Th	- entity which is moved or affected by some action (change of location or possession, object of transfer) - entity whose location is specified - test theme vs. agent: add "on purpose" to the sentence - if it does not fit, then it is (mostly) a theme, if it does fit, then agent - test theme vs. patient: does the referent change its quality during the action? – if no, then theme

Semantic Role	Abbreviation	Comment
		- object of possession (possessee)
Time	Time	- point or an interval of time

The following charts shows some examples of tagging Semantic Roles. In example (10), the theme (grammatical subject) is covertly realized.

(10)

ref	KAT_18880823_Copperman_flk.002 (002)		
tx	jukänət	än'šux	ɔɔlst.
mb	jukä-nət	än'šux	ɔɔl-s-t
mp	jukä-nət	än'šuk	ɔl-s-t
ge	woman-COM	old.man.[NOM.SG]	be-PST-3PL.S
SeR	np.h:Com	np.h:Th	v:pred
fe	There lived a woman and an old man.		

In example (11), the possessor is not realized covertl, thus it is tagged at the noun. The possessive suffix refers to the possessor.

(11)

ref	KAI_1904_TwoGirls_flk.010 (001.010)			
tx	ääwit'i	oone	kjeer	koort.
mb	ääw-i-t'-i	oon-e	kjeer	koor-t
mp	äw-ə-t'-ti	woon-i	kwiir	koor-ta
ge	girl-EP-DIM-3SG.[NOM]	sit-3SG.S	stove.[NOM.SG]	edge-LOC
SeR	n.h:A 0.3.h:Poss		n:Poss	n:Loc
fe	His daughter is sitting by the stove.			

In example (12), the subject *jükä* 'woman' is tagged as experiencer, since it depends on the emotional verb 'to see'. In the second clause, the subject is not expressed covertly. It is tagged as covert Theme at the predicate.

(12)

ref	KAI_1904_TwoGirls_flk.061 (001.061				
tx	jükä	aŋlɔtst,	kanɔw	pɛits	tɛrmälim.
mb	jükä	aŋlɔt-s-t	kanɔw	pɛit-s	tɛrmäl-im
mp	jükä	ɔŋlɔt-s-t	kɔnɔu	pät-s	tärmäl-əm
ge	woman.[NOM.SG]	see-PST-3SG.O	backward	fall-PST.[3SG.S]	get.frightened-CVB
SeR	np.h:E	0.3:Th	go:adv	0.3.h:Th	
fe	The children loved coming with together him.				

3.3.10.Information status (IST)

The Information status tier (IST) contains the annotation of information status. The annotation schema is based on the annotation guidelines for information structure and information status in Götze et al. (2007); the principles of annotation and the annotation scheme itself were developed by Wagner-Nagy et al. (2018: 28ff.) and made available by them. According to Götze et al. (2007: 150) the information status [a.k.a. activation, cognitive status, givenness] of a discourse referent reflects its retrievability within the discourse in question. A referent can be either given, accessible or new which can be determined by using the parameters [±discourse-old] and [±hearer-old]:

Table 10: Parameters for determining information status

	+discourse-old	**- discourse-old**
+hearer-old	given	accessible
- hearer-old	---	new

In detail, this means that given referents are necessarily and per default aforementioned in the discourse while accessible and new referents are not. Accessible referents can be somehow (see below) inferred by the "hearer" of the discourse. Hence, new referents are neither aforementioned nor inferable for the hearer. The basic tags for annotating information status are *giv*, *accs* and *new*, the extended tag set can be seen from the following table:

Table 11: Basic tags for annotating information status

Tag	**Comment**
Given referents	
giv-active	given and active referent (i.e. mentioned in the current or last sentence)
giv-inactive	given and inactive referent (i.e. mentioned before the last sentence)
Accessible referents	
accs-sit	referent, accessible through the situation (e.g. having breakfast: "Give me the butter, please.")
accs-aggr	referent, accessible through the aggregation of other referents (e.g. "*Unce upon a time, a king had a wife and two children. They lived happily.*")
accs-inf	referent, accessible through inference, e.g. part-whole relations (e.g. "*We had a turkey for thanksgiving. I ate its wings.*")
accs-gen	referent, accessible through general knowledge (e.g. "*The president of the U.S. travelled to Cuba.*")
New referents	
new	new referent

The corpus is not yet annotated for IST.

3.3.11.Borrowing (BOR)

The Borrowing tier (**BOR**) contains the annotation of borrowed lexical items. Both the origin of the item in question and the type of borrowing is annotated. The tags are made up as follows: <LANGUAGE:type>. The annotation is implemented already in the FLEx lexicon and automatically exported to EXMARaLDA. For Mansi there are Russian (RUS), Tatar (TAT), Iranina (IR), Khanty (KHAN), and Komi (KOMI). For the type of borrowing the following tags are used (cf. also Arkhipov (2020).

Table 12a: Tags for annotating borrowings

Tag	**Comment**
:cult	cultural borrowing (most frequent; also used for borrowed names)
:core	core borrowing
:gram	grammatical device (e.g. conjunctions)
:mod	modal words
:disc	discourse markers

Table 12b: Tags for annotating borrowings in the corpus

Tag	Description
Russian borrowing	
RUS:cult	lexical cultural borrowing from Russian
RUS:core	lexical core borrowing from Russian
%RUS:core	possible lexical core borrowing from Russian
RUS:gram	grammatical borrowing from Russian
RUS:disc	discourse word borrowing from Russian
Komi borrowing	

KOMI:cult	lexical cultural borrowing from komi
KOMI:core	lexical core borrowing from Komi
%KOMI:core	possible lexical core borrowing from Komi
Tatar borrowing	
TAT:cult	lexical cultural borrowing from Tatar
TAT:core	lexical core borrowing from Tatar
TAT:gram	grammatical borrowing from Tatar
Khanty borrowing	
KHAN:core	lexical core borrowing from Khanty
Other borrowings	
IR:core	lexical core borrowing from Iranian

Chart (13) shows the annotationof borrowings.

(13)

ref	MMA_19031220_FooledDevil2_flk.008 (001.008)			
tx	kompäälәŋ	jeent	šalkan	tolmantoŋ.
mp	kumpäälәŋ	jöu-nt	čɔlkɔn	tumlant-uŋ
ge	spirit.of.forest.[NOM.SG]	come-PRS.[3SG.S]	turnip.[NOM.SG]	steel-INF
BOR			TAT:cult	
fe	A forest spirit comes to steal turnips.			

3.3.12.Borrowing phonology, morphology and source

The tier **BOR-Phon** contains the annotation of phonological processes in borrowing. The tag set is the following.

Table 12 Annotation panel for phonological processes in borrowings

Tag	**Comment**
Deletions	
inCdel	initial consonant deletion
inVdel	initial vowel deletion (aphaeresis)
medCdel	medial consonant deletion
medVdel	medial vowel deletion (syncope)
finCdel	final consonant deletion
finVdel	final vowel deletion (apocope)
Insertions	
inVins	initial vowel insertion
medVins	medial vowel insertion
finVins	final vowel insertion
Substitutions	
Csub	consonant substitution
Vsub	vowel substitution
Other	
lenition	lenition (weakening)
fortition	fortition (strengthening)

The tier **BOR-Morph** contains the annotation of morphological processes in borrowing. The tags are made up as follows: <Strategy:Inflection>. The tag set is the following:

Table 13: Tags for annotating morphological processes in borrowings

Tag	**Comment**
Adaptation strategies	
dir:	direct insertion (i.e. insertion without morphological adaptation)
indir:	indirect insertion (i.e. insertion with morphological adaptation)

parad:	paradigm insertion (i.e. a paradigm borrowed)
Further inflection (in the matrix language)	
:bare	no inflection
:infl	further inflection

The tier BOR-Src gives the source form oft he borrowed word. The following chart shows some examples of annotating both borrowing phonology and borrowing morphology:

(14)

ref	MMA_19031220_FooledDevil2_flk.031 (001.031)		
tx	tawai,	länt,	šomew!
mp	tawaj	lää-nt	šom-ə-w
ge	let	say-PRS.[3SG.S]	run-EP-1PL.S
BOR	RUS:gram		
BOR-Phon	Csub		
BOR-Morph	parad:bare		
BOR-Src	давай		
fe	"Let's run," — he says.		

3.3.13.Code switching (CS)

The code-switching tier (CS) contains the code-switching annotation. Whereas borrowings deal with single words, code-switching deals (mostly) with sequences of two or more words. Both the language of the code-switch and the type of code-switch are annotated, using the <LANGUAGE:type> scheme. The language is mostly Russian (RUS).

Table 16: Tags for annotating code-switching

Tag	Comment
Sentence-external code-switching	
:ext	languages change at sentence (clause, utterance) borders
Sentence-internal code-switching	
:int.ins	languages change at phrase borders (e.g. a VP, NP, PP etc. is inserted)
:int.alt	the point of change is somewhere at an arbitrary point in the sentence

Code-switching is not yet annotated in this corpus.

3.3.14.Existential, locative and possessive predication (ExLocPoss)

The annotation schema was by Chris Lasse Däbritz. The ExLocPoss tier provides the annotation of existential, locative and possessive predication. Existential and locative predications express the temporary presence or absence of a referent X (theme) at a place Y (location). The theme is prototypically definite and topical in locative clauses, whereas it is prototypically indefinite in existential clauses. Possessive predications express that one referent Y (possessee) belongs to another referent X (possessor); prototypically, this relationship is again temporary, and the possessor has control over the possessee. In the case of inalienable possession (mostly, kinship and body terms), the latter does not hold.

The annotation scheme includes the three functional domains *existential* (Ex), *locative* (Loc) and *possessive* (Poss), the coding strategy, as well as the polarity (Aff or Neg) of the clause. The annotation tags have the format **Domain:Strategy.Polarity**. Table 14 shows the tags used for annotating existential, locative and possessive predication.

Table 14: Tags for annotating existential, locative and possessive predication

Tag	Comment
Functional domain	
Ex	existential predication
Loc	locative predication

Poss	possessive predication
Coding strategy and polarity	
:Zero.Aff	no lexical linking element, does not exclude pn-marking of theme/possessor at ground/possessee; affirmative
:Zero.Neg	no lexical linking element, does not exclude pn-marking of theme/possessor at ground/possessee; negative
:Cop.Aff	copula as linking element; affirmative
:Cop.Neg	copula as linking element; negative
:Ex.Aff	affirmative existential item as linking element
:Ex.Neg	negative existential item as linking element
:PosV.Aff	posture verb as linking element; affirmative
:PosV.Neg	posture verb as linking element; negative

The following charts show examples of the annotation of existential, locative and possessive clauses.

(15)

ref	MMA_19031220_FooledDevil2_flk.006 (001.006)			
tx	maa	paaxt	aano	toon’č’as.
mp	maa	pɔɔk-ta	ɔɔnu	toon’s’-ə-s
ge	meadow.[NOM.SG]	side-LOC	fir.[NOM.SG]	stand-EP-PST.[3SG.S]
ExLocPoss	Ex:PosV.Aff			
fe	There was a fir tree next to the field.			

3.3.15.Free translation (fe, fg, fh, fr)

The free translation tiers (**fe**, **fg, fh** and **fr**) give free translation of the utterance in question into English, German, Hungarian, and Russian. The translations are free, i.e. they do NOT necessarily reflect morphological and syntactical properties of the Mansi original. The translations follow the common guidelines presented in Arkhipov (2020). The following chart shows an example.

(16)

ref	MMA_19031220_FooledDevil2_flk.006 (001.006)			
tx	maa	paaxt	aano	toon’č’as.
mp	maa	pɔɔk-ta	ɔɔnu	toon’s’-ə-s
ge	meadow.[NOM.SG]	side-LOC	fir.[NOM.SG]	stand-EP-PST.[3SG.S]
fe	There was a fir tree next to the field.			
fg	Neben dem Acker stand eine Fichte.			
fh	A mezőnél egy fenyő állt.			
fr	В поле стояла ёлка.			

3.3.16.Literal German and Hugarian translation (ltg, lth)

The Literal German translation tier (**ltg**) contains the original German translation of the sentence in question. In case of the texts from Kannisto’s works, this means the published translation made by Kannisto and Liimola. In case of the texts taken from Munkácsi (1896), the Hungarian translations (**lth**) were made by Munkácsi.

(17)

ref	MMA_19031220_FooledDevil2_flk.006 (001.006)			
tx	maa	paaxt	aano	toon’č’as.
ge	meadow.[NOM.SG]	side-LOC	fir.[NOM.SG]	stand-EP-PST.[3SG.S]
fe	There was a fir tree next to the field.			
fg	Neben dem Acker stand eine Fichte.			
ltg	Neben dem Acker stand eine Föhre.			

3.3.17. Notes (nt)

The Notes tier (**nt**) eventually contains notes which clarify the content of the sentence or point at something peculiar in the sentence. The notes begin with the indication of who made the note (abbreviation as listed in 2.5.6, in square brackets, followed by a colon).

3.4. Searching the corpus

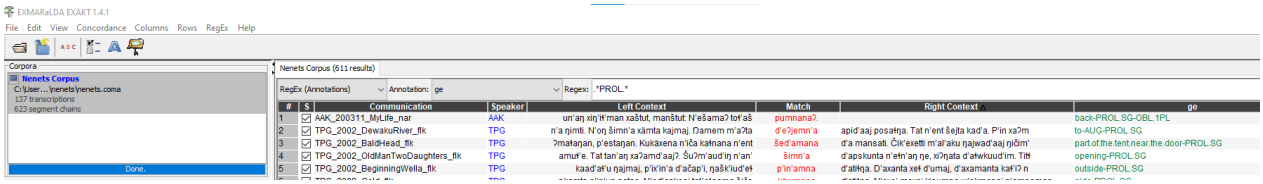
3.4.1. Search with EXMARaLDA EXAKT

The EXMARaLDA software suite includes EXAKT, an analysis and concordance tool.

In order to perform a search on the downloaded corpus files locally, the main metadata file (**tavda-mansi.coma**) should be opened with "File > Open Corpus" command. (Creating a word list is optional.)

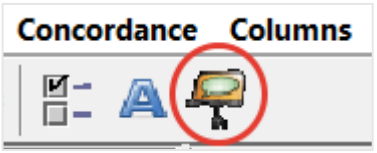
- One of the tiers should be selected in the main concordance window: either one of the annotation tiers (recommended; use "RegEx (Annotations)"; select any of tiers except **tx** under "Annotation") or the transcription tier (**tx**; use "RegEx (Transcription)").
- A search expression (interpreted as a regular expression[14]) should be specified in the **Regex** field. The matching results will be displayed in a column corresponding to the selected tier, e.g. "ge". Please refer to sections 3.2–3.3, Appendix 1 for annotations used in the corpus.
- Note that only the part matching the search expression will be displayed in the column. E.g. when searching for a prolative case marker with "PROL" in tier **ge**, only "PROL" will be shown in the "ge" column. in order to have the complete word gloss displayed in the "ge" column, enter ".*PROL.*" as search expression.

Figure 1. EXAKT search window



- The "Match" column represents the content of the **tx** tier (word or sentence) corresponding to the annotation found in the specified tier. Double-click the entry in the "Match" column to display a portion of the entire transcript containing the example found (all tiers) in the lower part of the screen. After that, a click on the "Open Partitur" button will open the entire transcript in EXMARaLDA Partitur Editor.

Figure 2.EXAKT: "Open Partitur" button



Please refer to EXMARaLDA manuals[15] for further details on using EXAKT and Partitur Editor.

3.4.2. Online search in Tsakorpus

Online search in the corpus is provided via Tsakorpus, an open-source search platform for linguistic corpora. The current version of the corpus can be accessed at https://inel.corpora.uni-hamburg.de/TavdaMansiCorpus/search. The interface of online search is available in English and in Russian.

Tsakorpus offers the following possibilities:

- Search in multiple annotation tiers
- Search for substring, simple patterns (using *) or regular expressions
- Multi-word search (with or without distance restrictions)
- Negative queries (sentences which do NOT have a word with specified parameters)

[14] https://www.regular-expressions.info/, last accessed 15.11.2024.
[15] https://exmaralda.org/en/quickstart-documents/, last accessed: 15.11.2024.

- Search for sentences, words (wordforms), lemmas
- Search in a subcorpus
- Exporting search results as CSV/XLSX

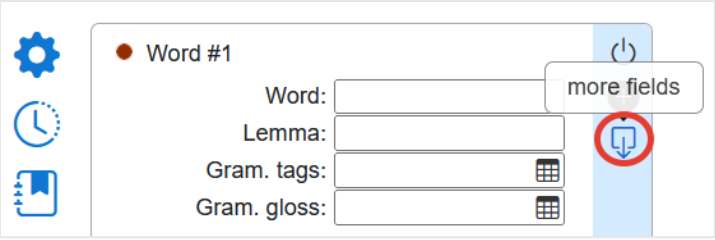
To run a search in the main transcription tier (**tx**) or in the word- and morph-level annotation tiers, "Language/tier" field should be set to "tavda Mansi" and the search expression(s) entered in one or more corresponding fields.

Table 13. Tsakorpus search fields and EXMARaLDA tiers: main transcription and word-/morph-level annotation

Tsakorpus search field	**Corresponding tier in EXMARaLDA**
Word	**tx**
Lemma	**mp** (stem)
Gram. tags	**ps**; grammar tags generated from grammatical glosses (**ge, gr**)
Gram. gloss	grammatical (i.e. affix) glosses (**ge, gr**)
Lexical gloss (en)*	lexical (i.e. stem) glosses (**ge**)
Lexical gloss (ru)*	lexical (i.e. stem) glosses (**gr**)
Morph. slot*	**mc**
Semantic role*	**SeR**
Syntactic function*	**SyF**
Inform. status*	**IST**
Borrowing*	**BOR**
Bor. phonetics*	**BOR-Phon**
Bor. morphology*	**BOR-Morph**
Bor. source*	**BOR-Src**
Code-switching*	**CS**
Geogr. coordinates*	**geo**
Exist/loc/poss*	**ExLocPoss**

*To display search fields marked with *, click on "More fields" button next to "Word" and "Lemma" fields.

Figure 3. Tsakorpus interface: Show more fields



Lexical and grammatical glosses in Tsakorpus

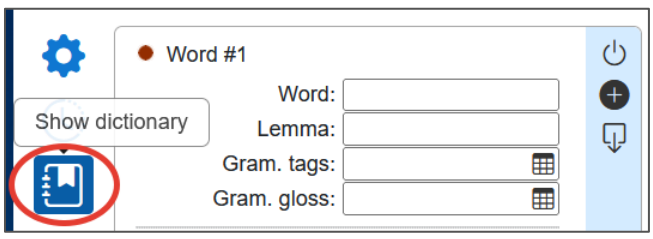
Each word in Tsakorpus is internally split into stems (lexical items) and affixes (grammatical morphs).

The stem can be found by searching for its underlying (**mp**) form (e.g. "iis") in the **Lemma** field, or by searching for its lexical gloss (e.g. "brain" / "ум") in Lex. gloss (en) or Lex. gloss (ru) fields.

The affixes can be found by searching for the complete gloss (e.g. "NOM.SG") in the **Gram. gloss** field, or with corresponding grammar tags (e.g. "nom,sg") in the **Gram. tags** field (see next section for details on grammar tags).

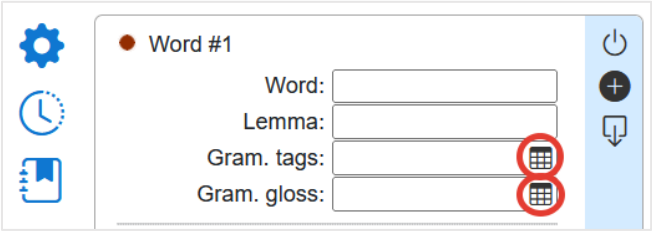
A list of lemmas (i.e. underlying forms of stems as represented in **mp** tier) along with their translations (lexical glosses) can be displayed with "Show dictionary" button.

Figure 4. Tsakorpus interface: Show dictionary



For most word- and morph-level annotation tiers, such as grammar tags, grammatical glosses, borrowings, one can either type in the search expression directly or choose from the list of available values. To open the list of values, click on the icon in the search field.

Figure 5. Tsakorpus interface: Show list of values



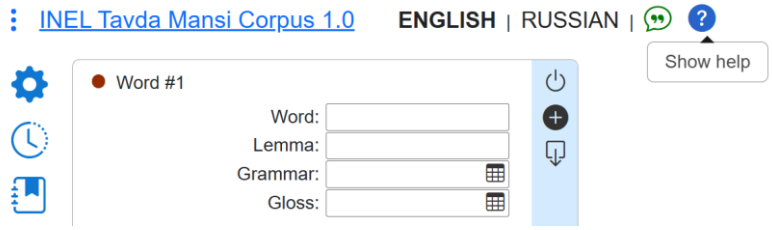
Grammatical glosses and grammar tags in Tsakorpus

In addition to grammatical glosses as present in tiers **ge**, **gr**, Tsakorpus provides another search possibility called "grammar tags." Grammar tags are generated by rules based on part of speech and glosses. For a complete list of glosses and grammar tags please refer to Appendix 1.

- Tags are assigned to an entire word and not to a particular morpheme in a word.
- By default, grammar tags are identical to a lower-case version of the corresponding gloss or part of speech label, e.g. (part of speech) "v" => "v", (gloss) "DU" => "du", "ADJZ" => "adjz". Exceptions are mostly due to avoiding overlapping.
- Parts of speech can only be found with grammar tags since they do not have a corresponding gloss.
- Stems with glossing labels similar to a grammatical gloss, e.g. "NEG" for "negative verb" (see previous section), will also be assigned grammar tags. Such glosses are marked as "lexical" in Comments columns in Appendix 1. They can be found with either **Gram. tags** or **Lex. gloss (en)** / **Lex. gloss (ru)** fields, but not with **Gram. gloss** field.
- A group of related glosses can get more than one tag each to allow different ways of searching. E.g. of the two hortative markers, "HORT1" will get tags "hort,hort1" and "HORT2" will get tags "hort,hort2". Therefore each of them separately can be found with their distinctive tags (resp. "hort1" and "hort2"), while searching for "hort" will find both of them. Please refer to Appendix 1 for complete lists of tags.
- Zero morphs have no overt segment in **mb**, **mp** tiers, and their glosses are shown in square brackets preceded by a dot in **ge**, **gr** tiers. In Tsakorpus, they can only be found with corresponding grammar tags.
- When specifying more than one tag in a search expression, they can be combined with logical operators: AND (","), OR ("|") and NOT ("~"), e.g. "v,inch,md" or "n,(abl|loc)". When selecting tags from the list of values, multiple tags which are listed as belonging to the same Tsakorpus category (see Appendix 1) will be by default joined by OR ("|"), e.g. "(abl|loc)". Multiple tags which are listed as belonging to different Tsakorpus categories will be by default joined by AND (","), e.g. "v,inch,md".

For further details please refer to Tsakorpus online help.

Figure 6. Tsakorpus interface: Show help



References

Arkhipov, A. 2020: *INEL Corpora General Transcription and Annotation Principles*. Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology. Vol. 5. Szeged; Hamburg. DOI: 10.14232/wpcl.2020.5

Arkhipov Alexandre & Chris Lasse Däbritz (2018): Hamburg corpora for indigenous Northern Eurasian languages. *Tomsk Journal of Linguistics and Anthropology*. Issue 3 (21): 9–18. URL: https://ling.tspu.edu.ru/en/archive.html?year=2018&issue=3&article_id=7130

GOST 2001 — *GOST 7.79 – 2000. Sistema standartov po informacii, bibliotečnomu i izdatel'skomu delu. Pravila transliteracii kirillovskogo pis'ma latinskim alfavitom* [GOST 7.79 – 2000. The system of standards in information, librarianship and publishing. Rules of transliteration of Cyrillic letters into the Latin alphabet]. Minsk: Mežgosudarstvennyj sovet po standartizacii, metrologii I sertifikacii. https://protect.gost.ru/document.aspx?control=7&id=130715, https://ifap.ru/library/gost/7792000.pdf

Götze, Michael, Thomas Weskott, Cornelia Endriss, Ines Fiedler, Stefan Hinterwimmer, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas, Ruben Stoel 2007: Information structure. In Dipper, S., M. Götze, Stavros Skopeteas (eds): *Information Structure in Cross-Linguistic Corpora* (*Interdisciplinary Studies on Information Structure* 07, pp. 147–187). [URL: https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/deliver/index/docId/2036/file/Kapitel6_07.pdf [Accessed: 11.04.2025]

Haig, Geofry, & Schnell, Stefan 2014: Annotations using GRAID (Grammatical relations and animacy in discourse). *Introduction and guidelines for annotators, Version* 7.0. URL: https://multicast.aspra.uni-bamberg.de/data/pubs/graid/Haig+Schnell2014_GRAID-manual_v7.0.pdf [Accessed: 01.04.2025]

Honti, László 1975: *System der paradigmatischen Suffixmorpheme des wogulischen Dialektes an der Tawda*. den Haag/Parin: Mounton.

Kannisto, Artturi 1907: Matkakertomus vogulimailta, III. *Suomalais-ugrilaisen seuran aikakauskirja* 24: 1–3.

Kannisto, Artturi & Liimola, Matti 1951: *Wogulische Volksdichtung* gesammelt und übersetzt von Artturi Kannisto, bearbeitet und herausgegeben von Matti Liimola. Band I. Texte mythischen Inhalts. Helsinki: Suomalais-Ugrilainen Seura.

Kannisto, Artturi & Liimola, Matti 1956: Wogulische Volksdichtung gesammelt und übersetzt von Artturi Kannisto, bearbeitet und herausgegeben von Matti Liimola. Band VI. Märchen. Helsinki: Suomalais-Ugrilainen Seura.

Munkácsi, Bernát 1894: *A vogul nyelvjárasok szóragozásukban ismertetve.* [Inflection in vogul dialects]. Budapest: Magyar Tudományos Akademia.

Munkácsi, Bernát 1896: *Vogul népköltési gyűjtemény* [Vogul folklore collection] Volume IV. Budapest: Magyar Tudományos Akadémia.

VPN 2010 = *Vserossijskaya perepis` naseleniya 2010. Tom 4. Nacional`ny`j sostav i vladenie yazy`kami.* Available online at: http://www.gks.ru/free_doc/new_site/perepis2010/croc/Documents/Vol4/pub-04-05.pdf. [Accessed: 19.08.2019].

Wagner-Nagy, Beáta & Szeverényi, Sándor & Gusev, Valentin. 2018. *User's Guide to Nganasan Spoken Language Corpus.* Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology 1. University of Szeged, Department of Finno-Ugric Studies & Universität Hamburg, Zentrum für Sprachkorpora: Szeged, Hamburg. https://doi.org/10.14232/wpcl.2018.1

Literature on Tavda Mansi

Bakró-Nagy Marianne: 19th century fieldwork methodology in Southern Mansi and the reliability of data. https://www.babel.gwi.uni-muenchen.de/media/downloads/bakro-nagy_leiden_presentation.pdf

Honti, László (1969): A tárgy jelölése a vogul nyelv tavdai nyelvjárásában. *Nyelvtudományi Közlemények* 71: 113-120.

Honti László (1975): System der paradigmatischen Suffixmorphemen des wogulischen Dialektes an der Tawda. Budapest: Akadémiai Kiadó

Honti, László (1980): Zur Phonemanalyse des Tavda-Wogulischen. Finnisch-Ugrische Mitteilungen 4: 61-88.

Munkácsi, Bernát (1894): *A vogul nyelvjárások szóragozásukban ismertetve*. Budapest: Magyar Tudományos Akadémia.

Munkácsi, Bernát & Béla Kálmán (1986): *Wogulisches Wörterbuch*. Budapest: Akadémiai Kiadó.

Sherwood, Peter (1996): Megjegyzések a tavdai vogul végnapjairól. [Notes on the end days of Tavda Vogul] In: Leskinen, Heikki, Sándor Maticsák & Tõnu Seilenthal (eds.): Congressus Octavus Internationalis Fenno-Ugristarum (CIFU VIII) Pars IV. 1995–1996 Gummerus, Jyväskylä. 244-248.

Szilágyi Norbert (2012): A déli manysi nyelvjárás tárgyjelölésének szintaktikai szempontú vizsgálata. In Gécseg, Zsuzsa (ed): Lingdok 14. Nyelvészdoktoranduszok dolgozata. Szeged, 199–220. http://nydi.szte.hu/LingDok_konfkotetek_files/lingdok14.pdf

Appendix 1. Morpheme glossing labels (ge, gr)

Tag	Description	Tsakorpus grammar tags	Tsakorpus category	Gloss type
Person and number				
1SG	1 person singular	pn1,pnsg	persnum-person,persnum-number	
1SG.EMPH	1 person singular, emphatic marker	pn1,pnsg,emph	persnum-person,persnum-number,disc	
1SG.S	1 person singular, subjective conjugation	pn1,pnsg,subjc	persnum-person,persnum-number,persnum-conj	
1SG.O	1 person singular, objective conjugation	pn1,pnsg,objc	persnum-person,persnum-number,persnum-conj	
1SG.OPL	1 person singular, objective conjugation, plural object	pn1,pnsg,objc,objpl	persnum-person,persnum-number,persnum-conj,persnum-num	
1SG.ACC	1 person singular, accusative case	pn1,pnsg,acc	persnum-person,persnum-number,case	
1SG.DAT	1 person singular, dative case	pn1,pnsg,dat	persnum-person,persnum-number,case	
1PL	1 person plural	pn1,pnpl	persnum-person,persnum-number	
1PL.S	1 person plural, subjective conjugation	pn1,pnpl,subjc	persnum-person,persnum-number,persnum-conj	
1PL.O	1 person plural, objective conjugation	pn1,pnpl,objc	persnum-person,persnum-number,persnum-conj	
1PL.OPL	1 person plural, objective conjugation, plural object	pn1,pnpl,objc,objpl	persnum-person,persnum-number,persnum-conj,persnum-num	
1PL.ACC	1 person plural, accusative case	pn1,pnpl,acc	persnum-person,persnum-number,case	
1PL.DAT	1 person plural, dative case	pn1,pnpl,dat	persnum-person,persnum-number,case	
2SG	2 person singular	pn2,pnsg	persnum-person,persnum-number	
2SG.EMPH	2 person singular, emphatic marker	pn2,pnsg,emph	persnum-person,persnum-number,disc	

2SG.S	2 person singular, subjective conjugation	pn2,pnsg,subjc	persnum-person,persnum-number,persnum-conj	
2SG.O	2 person singular, objective conjugation	pn2,pnsg,objc	persnum-person,persnum-number,persnum-conj	
2SG.ACC	2 person singular, accusative case	pn2,pnsg,acc	persnum-person,persnum-number,case	
2SG.DAT	2 person singular, dative case	pn2,pnsg,dat	persnum-person,persnum-number,case	
2SG.POSS	2 person singular, possessive	pn2,pnsg,poss	persnum-person,persnum-number,poss	
2DU	2 person dual	pn2,pndu	persnum-person,persnum-number	only as a tag
2PL	2 person plural	pn2,pnpl	persnum-person,persnum-number	
2PL.S	2 person plural, subjective conjugation	pn2,pnpl,subjc	persnum-person,persnum-number,persnum-conj	
2PL.O	2 person plural, objective conjugation	pn2,pnpl,objc	persnum-person,persnum-number,persnum-conj	
2PL.ACC	2 person plural, accusative case	pn2,pnpl,acc	persnum-person,persnum-number,case	
3SG	3 person singular	pn3,pnsg	persnum-person,persnum-number	
3SG.EMPH	3 person singular, emphatic marker	pn3,pnsg,emph	persnum-person,persnum-number,disc	
3SG.S	3 person singular, subjective conjugation	pn3,pnsg,subjc	persnum-person,persnum-number,persnum-conj	
3SG.O	3 person singular, objective conjugation	pn3,pnsg,objc	persnum-person,persnum-number,persnum-conj	
3SG.OPL	3 person singular, objective conjugation, plural object	pn3,pnsg,objc,objpl	persnum-person,persnum-number,persnum-conj,persnum-num	
3SG.DAT	3 person singular, dative case	pn3,pnsg,dat	persnum-person,persnum-number,case	

3PL	3 person plural	pn3,pnpl	persnum-person,persnum-number	
3PL.S	3 person plural, subjective conjugation	pn3,pnpl,subjc	persnum-person,persnum-number,persnum-conj	
3PL.O	3 person plural, objective conjugation	pn3,pnpl,objc	persnum-person,persnum-number,persnum-conj	
3PL.OPL	3 person plural, objective conjugation, plural object	pn3,pnpl,objc,objpl	persnum-person,persnum-number,persnum-conj,persnum-num	
Conjugation type				
S	subjective conjugation	subjc	persnum-conj	only as a tag
O	objective conjugation	objc	persnum-conj	only as a tag
OPL	objective conjugation, plural object	objc,objpl	persnum-conj,persnum-num	only as a tag
MD	middle conjugation	md	persnum-conj	
Personal pronouns				
PRO1SG	personal pronoun, 1 person singular	pn1,pnsg	persnum-person,persnum-number	lexical
PRO1DU	personal pronoun, 1 person dual	pn1,pndu	persnum-person,persnum-number	lexical
PRO1PL	personal pronoun, 1 person plural	pn1,pnpl	persnum-person,persnum-number	lexical
PRO2SG	personal pronoun, 2 person singular	pn2,pnsg	persnum-person,persnum-number	lexical
PRO2DU	personal pronoun, 2 person dual	pn2,pndu	persnum-person,persnum-number	lexical
PRO2PL	personal pronoun, 2 person plural	pn2,pnpl	persnum-person,persnum-number	lexical
PRO3SG	personal pronoun, 3 person singular	pn3,pnsg	persnum-person,persnum-number	lexical
PRO3PL	personal pronoun, 3 person plural	pn3,pnpl	persnum-person,persnum-number	lexical
Nominal categories				
Number				
SG	singular number	sg	num	only as a tag
SG.1SG	singular number, 1 person singular	sg,poss,pn1,pnsg	num,persnum-person,persnum-number	

SG.1PL	singular number, 1 person plural	sg,poss,pn1,pnpl	num,persnum-person,persnum-number	
SG.2SG	singular number, 2 person singular	sg,poss,pn2,pnsg	num,persnum-person,persnum-number	
SG.2DU	singular number, 2 person dual	sg,poss,pn2,pndu	num,persnum-person,persnum-number	
SG.2PL	singular number, 2 person plural	sg,poss,pn2,pnpl	num,persnum-person,persnum-number	
SG.3SG	singular number, 3 person singular	sg,poss,pn3,pnsg	num,persnum-person,persnum-number	
SG.3SG.S	singular number, subjective conjugation	sg,poss,pn3,pnsg,subjc	num,persnum-person,persnum-number,persnum-conj	
SG.3PL	singular number, 3 person plural	sg,poss,pn3,pnpl	num,persnum-person,persnum-number	
PL	plural number	pl	num	
PL.1SG	plural number, 1 person singular	pl,poss,pn1,pnsg	num,persnum-person,persnum-number	
PL.2SG	plural number, 2 person singular	pl,poss,pn2,pnsg	num,persnum-person,persnum-number	
PL.2PL	plural number, 2 person plural	pl,poss,pn2,pnpl	num,persnum-person,persnum-number	
PL.3SG	plural number, 3 person singular	pl,poss,pn3,pnsg	num,persnum-person,persnum-number	
PL.3PL	plural number, 3 person plural	pl,poss,pn3,pnpl	num,persnum-person,persnum-number	
Case				
ABL	ablative case	abl	case	
ADV	adverbial case	adv	case	
ADV.LAT	adverbial lative case	adv,lat	case,case	
ACC	accusative case	acc	case	
COM	comitative case	com	case	
DAT	dative case	dat	case	only as a tag
LAT	lative case	lat	case	
LOC	locative case	loc	case	
NOM	nominative case	nom	case	
NOM.SG	nominative case, singular number	nom,sg	case,num	
OBL	oblique case	obl	case	only as a tag
Possession				
POSS	possessive	poss	poss	only as a tag
Other nominal categories				

INSTR	instrumental nominal	instr	deriv-n	
NMLZ	nominalization	nmlz	deriv-n	
ABSTR	abstract suffix	abstr	cat	
COLL	collective suffix	coll	cat	
Numerals				
ORD	ordinal numeral	ord	deriv-num	
DYA	dyadic (connective-reciprocal)	dya	deriv-num	
Verbal categories				
TAM categories				
FRQ	frequentative	frq	tam	
HORT	hortative	hort	tam	
JUS	jussive	jus	tam	
IMP	imperative	imp	tam	
INCH	inchoative	inch	tam	
MOM	momentaneous	mom	tam	
PRS	present tense	prs	tam	
PST	past tense	pst	tam	
PRF	perfective	prf	tam	
Non-finite forms				
COND	conditional converb	cond	nfin	
CVB	converb	cvb	nfin	
INF	infinitive	inf	nfin	
PTCP	participle	ptcp	nfin	only as a tag
PTCP.PST	past participle	ptcp,pst	nfin,tam	
PTCP.PRS	present participle	ptcp,prs	nfin,tam	
PTCP.ABES	abessive participle	ptcp,abes	nfin,nfin	
Other verbal categories				
CAUS	causative	caus	vrb	
DRV	unspecified derivation	drv	vrb	
EP	epenthetic element	ep	misc	
PASS	passive participle	pass	vrb	
ITER	iterativee	iter	vrb	
RFL	reflexive	rfl	vrb	
VBLZ	verbalizer	vblz	cat	
Negation				
NEG	negation	neg	neg	lexical
NEG.IMP	negative imperative	neg,imp	neg,neg	lexical
NEG.EX	existential negation	neg,ex	neg,neg	lexical
Other categories				
ADJZ	adjectivizer	adjz	cat	
CAR	caritive	car	cat	
COMP	comparative	comp	comp	
DIM	diminutive (any)	dim	deriv-n	
TRL	translative	trl	cat	
EMPH	emphatic marker	emph	disc	
INDF	indefinite marker	indf	disc	

INTERJ	unspecified interjection	interj	disc	
%%	unknown morph	unkn	misc	

Appendix 2. Tavda Mansi morphemes in alphabetical order[16]

Marker	Abbreviation	Function
-Ø (zero)	NOM; NOM.SG	nominative case
	[3SG.S]	3sg verbal ending, subjective conjugation
-ai	DRV	derivative suffix
äx-	APPR.	approximative prefix
ɔlɔ-	INDF	indefinite prefix
-än	PL.3PL	possessive marker for 3PL possessor and PL possessum
	SG.3PL	possessive marker for 3PL possessor and SG possessum
-ə	EP	ephentetic vowel
-əm	CVB	converb suffix
-ännä	PL.2PL	possessive marker for 2PL possessor and PL possessum
-ət	ORD	ordinal numeral suffix
il-	PRF	perfectivizer preverb
-iin	2PL.S	2PL subjective conjugation
	IMP.2PL.S	imperative 2PL subjective conjugation
-iŋ	ADJZ	adjectivizer
	DRV	derivation suffix
-k	EMPH	emphatic suffix
-kä	COND	conditional mood
-käl	ITER	iterative suffix
-kət	MED	medial suffix
-kəu	ADV	adverbializer (Tatar dative (-ka/-kä) plus Mansi translative)
-kɛ	DRV	derivation suffix
-ki	DRV	derivation suffix (deverbal verb)
-kil	ADV	adverbial numeral marker
-l	INSTR	instrumental case
	VBLZ	verbalizer
	FRQ	frequentative suffix
-lɔt	DRV	derivation suffix (deverbal verb)
-läl	FRQ	frequentative suffix
-länəm	1SG.OPL	1SG verbal ending, objective conjugation plural object
-länu	1PL.OPL	1PL verbal ending, objective conjugation plural object
läp-	PRF	perfectivizer preverb
-lɔn	3PL.O	3PL verbal ending, objective conjugation, singular object
	3PL.OPL	3PL verbal ending, objective conjugation, plural object
-lək	ABSTR	abstract suffix
-ləm	1SG.O	1SG verbal ending, objective conjugation, singular object
-lən	2SG.O	2SG verbal ending, objective conjugation, singular object
-lənä	IMP.2PL.OPL	imperative 2PL objective conjugation plural object
-lm	MOM	momentan suffix
-lo	1PL.O	1PL verbal ending, objective conjugation singular object
-l'i	INSTR	instrumental noun suffix
-lt	CAUS	causative suffix
-m	MOM	momentan suffix
	SG.1SG	possessive marker for 1SG possessor and SG possessum
	PTCP.PST	past participle
	TRL	translative suffix (denominal verb)
-mäi	DRV	derivation suffix
-mät	DRV	derivation suffix (deverbal verb)
-mət	INCH	inchoative suffix

[16] Here, only these morphemes are listed, whose function is known. Morphemes glossed with <%%>, since their function is unknown, are not included.

-mej	DRV	derivation suffix
-men	SG.1DU	possessive marker for 1DU possessor and SG possessum
-mə	ACC	accusative case
-n	SG.2SG	possessive marker for 2SG possessor and SG possessum
	COLL	collective suffix
-nɔ	LAT	lative case
-na	2PL.S	2PL verbal ending, subjective conjugation
	SG.2PL	possessive marker for 2PL possessor and SG possessum
-nal	ABL	ablative case
-nəm	PL.1SG	possessive marker for 1SG possessor and PL possessum
-nət	COM	comitative case
-ne	PTCP.PRS	present participle
-nen	PL.2PL	possessive marker for 2PL possessor and PL possessum
-nɛ	COND	conditional mood
-ni	PL.3SG	possessive marker for 3SG possessor and PL possessum
-nt	FRQ	frequentative suffix
-n’s’i	DYA	dyadic (connective-reciproc) suffix
-ŋ	ADJZ	adjectivizer
-p	ADJZ	adjectivizer
	INSTR	instrumental noun suffix
	MOM	momentan suffix
	PTCP.PRS	present participle
-pt	CAUS	causative suffix
-s	DRV	derivation suffix (deverbal verb)
	FRQ	frequentative suffix
-skä	COND	conditional
-t	DRV	derivation suffix (deverbal verb)
	CAUS	causative suffix
	NMLZ	nominalizer
	PL	plural
	VBLZ	verbalizer
-täni	3SG.OPL	3SG verbal ending, objective conjugation, plural object
-təl	CAR	caritive
	PTCP.ABES	abessive participle
-tənnɔ	2PL.O	2PL verbal ending, objective conjugation, singular object
-tənəm	PL.1SG	possessive marker for 1SG possessor and PL possessum
-tk	DRV	derivative suffix
-ti	SG.3SG	possessive marker for 3SG possessor and SG possessum
-ta	LOC	locative case
-u	SG.1PL	possessive marker for 1pl possessor and sg possessum
	TRL	translative case
-uŋ	INF	infinitive
-w	1PL.S	1PL verbal ending, subjective conjugation