INEL Nenets corpus

User documentation

Josefina Budzisch, Beáta Wagner-Nagy, Alexandre Arkhipov, 2024

1. Introduction

1.1. Objective of the corpus

The present corpus of the Nenets language(s) has been developed as part of the long-term research project INEL ("Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages").

The corpus makes possible typologically aware corpus-based grammatical research on the Nenets language(s) and expands the documentation of the lesser described indigenous languages of Northern Eurasia.

1.2. Nenets language(s)

1.2.1. Description

Nenets belongs to the Samoyedic branch of the Uralic language family. Nenets is spoken from the Kanin Peninsula in Europe to the Taimyr Peninsula in Northwestern Siberia, including the Yamal-Nenets and Khanty-Mansi Autonomous Regions.

Nenets is at present endangered, and according to the census 2020, there are 49,646 people identifying themselves as Nenets and 24,487 self-identified speakers.

Nenets is an agglutinating-synthetic language, with its morphology characterized by a wide variety of inflectional and derivational suffixes (especially in verbs). The main declensional categories of Nenets nouns are number, case, possession, while verbs carry markers of several aspectual categories, mood and evidentiality, tense, as well as person and number agreement.

There are two Nenets lects: Tundra Nenets (NenT) and Forest Nenets (NenF), which can be considered mutually intelligible, but have a number of clear distinctions in lexicon, phonology, and morphology. They were traditionally presented as two dialects of the same language, but recently have been more and more frequently considered two different languages.

1.2.2. Language Codes

ISO-639-3 code: **yrk**

Glottolog code: **nene1249** for Tundra Nenets and **fore1274** for Forest Nenets

1.2.3. Dialectal subdivisions

Tundra Nenets has three dialects: western, central, and eastern, with subdialects in the western and eastern groups. Forest Nenets exhibits significant dialectal differences due to contact with neighboring peoples. Verbov (1973: 124) identified three main dialects: Pur, Lyamin, and Nyamin, with distinct linguistic features. Koshkareva (2005) notes changes in the 20th century, now identifying Pur, Agan, and Numto as the main dialects. We adopted this revised classification for our analysis.

Table 1. Nenets dialects (based on Koskareva 2005, Burkova 2010: 182 and Burkova 2022: 675)

Variety	Dialectgroup	Dialect	Territory
Tundra	Western	**Kanin**	Kanin tundra (from Kanin Peninsula to the river Pyosha)
		Timan	Timan tundra (from river Pyosha to river Indiga)
		Kolguyev	Kolguyev Island
		Malaya Zemlya	Malozemelskaya tundra, from river Indiga to river Pechora
	Central	**Bolshaya Zemlya**	Eastern part of the Nenets Autonomous District of the Arkhangelsk Oblast
	Eastern	Priural	From the eastern slopes of the Ural Mountains to the Yamal Peninsula
		Yamal	Yamal Peninsula
		Nadym	Basin of the river Nadym, western part of the Tazovsky Peninsula
		Taz	Northeastern part of the Tazovsky Peninsula, right bank of the Taz Bay, lower reaches of the river Taz
		Gyda	Western part of Gyda Peninsula
		Taimyr (Yenisey)	Western part of the Taimyr Peninsula and the basin of the lower river Yenisey
Forest		**Pur**	Upper, middle, and partially lower reaches of basin of the river Pur
		Agan	along the rivers Agan and Amputa
		Numto	area of lake Numto and the upper reaches of the river Kazym

The dialects represented in the corpus are highlighted in bold in Table 1.

1.3. Archiving

The corpus comprises source media files (whenever available), annotated transcripts in *EXMARaLDA*[1] transcript formats and metadata descriptions in *EXMARaLDA* Coma format (see 2.6 and 2.8 for details).

The corpus is archived and published by the Research Data Repository of the University of Hamburg[2] under open-access conditions with Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0).[3]

The corpus is available for download in three packages of different size (see 2.5.2 for details):

- The “standard” package includes sound files in WAV format.
- The “mp3” package includes sound files in MP3 format.
- The “lite” package does not include any sound or PDF files.

Besides the downloadable packages, the corpus is accessible online through Tsakorpus,[4] 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/NenetsCorpus/search.

[1] http://exmaralda.org/en/, last access: 07.06.2024.
[2] https://www.fdr.uni-hamburg.de/communities/inel, last access: 15.07.2024.
[3] https://creativecommons.org/licenses/by-nc-sa/4.0/, last access: 15.07.2024.
[4] https://github.com/timarkh/tsakorpus, last accessed: 17.11.2024.

1.4. Citation

The corpus is to be cited as follows:

Budzisch, Josefina; Wagner-Nagy, Beáta. 2024. INEL Nenets Corpus. Version 1.0. Publication date 2024-12-31. https://hdl.handle.net/11022/0000-0007-FE37-E. 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 Nenets 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.[5]

1.5. Project members

Project summary information

The INEL Nenets corpus has been created within the long-term INEL project ("Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages"), 2016–2033. For an overview of the project, see (Arkhipov & Däbritz 2018). The project homepage can be visited at: https://inel.corpora.uni-hamburg.de/.

The Nenets subproject spanned three years from January 2022 to December 2024. The research was carried out at the Institute for Finno-Ugric/Uralic Studies (IFUU) of the Hamburg University (UHH).

Project leader

Beáta Wagner-Nagy

Editors

Beáta Wagner-Nagy, Alexandre Arkhipov

Main corpus authors

Josefina Budzisch

Beáta Wagner-Nagy

Contributions of Nenets speakers and of particular researchers are acknowledged in more detail in the metadata to the corpus (see 2.8).

Technical developers

Timm Lehmberg (technical coordinator; until January 2023)

Aleksandr Riaposov

Elena Lazarenko

Student assistants

Natalia Kim (September 2023 – September 2024)

Symbat Saldyr (August 2023 – December 2024)

1.6. Acknowledgements

Funding

This corpus has been produced in 2022–2024 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

[5] As Budzisch 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.

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 Beáta Wagner-Nagy, Michael Rießler, Hanna Hedeland, and Timm Lehmberg.

Legacy data

The majority of texts in this corpus originate from published works, which are appropriately cited in the relevant sections of the metadata. In particular, the following publications were used, the full information can be found in the reference section:

- Barmich 2018
- Burkova 2008
- Burkova 2012
- Burkova et al. 2003
- Hajdú 1968
- Koshkareva et al. 2007
- Labanauskas 2001
- Logany & Logany 2016
- Lyubinskaya 2022
- Pusztay 1976
- Tereshchenko 1956
- Tereshchenko 1990
- Turutina 2003
- Yangasova 2018

Svetlana Burkova kindly shared a collection of her Forest Nenets data including an original sound recording (Agan dialect), transcripts and glosses as Toolbox files and Word documents (Agan and Pur dialects), as well as published texts in Pur (Turutina 2003) and Numto (Logany & Logany 2016) dialects.

2. The corpus

2.1. The language(s) of the corpus

2.1.1. Content

The language of content in the corpus is almost exclusively Nenets, in instances of code-switching Russian (part of them are left untranscribed in the corpus).

There is always only one main transcription tier (per speaker), using the common INEL transcription style (see 3.2.3).

2.1.2. Annotations

The main annotation language in the corpus is English.

The main content transcript is translated into English, Russian and German (see tiers **fe, fr, fg**), some texts were published with a Hungarian translation, these are also included for respective texts (see tier **fh**).

For texts from the written materials, original translation into Russian is given as provided in the published source (see tier **ltr**). For texts transcribed from the audio tapes, translation provided by the native speakers during transcription sessions is given in the same tier.

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 main language of the metadata is English. Russian spellings of personal names and place names are also provided. On transliteration of names, see 2.5.6.

2.2. Sources

The INEL Nenets corpus originates from published sources and materials that were provided by Svetlana Burkova stemming from her own fieldwork.

2.3. Content

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

2.4. Corpus size

Table 22 contains the summary of the number of speakers, number of texts ("communications," in EXMARaLDA terms) in total and specified by genre, number of sentences and number of tokens, separately for NenF and NenT subcorpora and in total.

Table 2. Corpus size

		NenF	**NenT**	**Total**
Speakers		**22**	**31**	**53**
Texts ("Communications")	Elicitation	0	1	1
	Folklore	56	39	95
	Narrative	24	15	39
	Song	0	1	1
	Total	**80**	**56**	**136**
Sentences		**3,709**	**6,545**	**10,254**
Tokens		**23,597**	**37,681**	**61,278**

Total duration of audio: ca. **45 min**

2.5. Naming Conventions

2.5.1. Name of the corpus

The name of the corpus is INEL Nenets Corpus.

2.5.2. Folder structure and filenames

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

Folders with text transcripts, organized by genre:

- “el” (elicited texts)
- “flk” (folklore texts)
- “nar” (narrative texts)
- “sng” (songs)

Each of these genre folders contains one further subfolder per text (“communication”), named identically to the text name (see 2.5.3). 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)
- (if available) sound file with the recording in WAV format (*.wav) [“standard” package] or MP3 format (*.mp3) [“mp3” package]
- (if available) original version coming from published or archive sources in PDF (*.pdf) format (scanned images or text depending on the available source; transcriptions originally written down as MS Word files and handwritten transcriptions are also kept here) [“standard” and “mp3” packages]
- Annotated transcripts and original audio files have the file names identical to the text name (see 2.5.3), except for the “_s” suffix.
- In the “lite" package of the corpus, no sound files are included.

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 3.3.7–3.3.12)
- "gloss-panel-nenets.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

Individual files:

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

2.5.3. 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 2.5.4); further speaker codes (optional); year of recording; short title; genre abbreviation. These components are joined by underscore ("_").

If the year of recording is only approximate or altogether unknown, a placeholder character "X" is used to fill the missing digits (e.g., "199X"). In the text metadata, more details on the date of recording can be specified if known (e.g., “early 1990s” instead of "199X").

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 2.5.5.

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

Communication code: YaVD_1976_OldKhantysSon_flk

Speaker code: YaVD (Yar, Vasilij Dmitrievich)

Date of recording: 1976

Short title: OldKhantysSon (i.e. “Old Khanty’s Son”)

Genre: flk (folklore)

2.5.4. 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. “PVL” stands for Pyak, Vasilij Lemevich (Пяк, Василий Лемевич). Still, letter combinations Ch, Sh, Yu, Ya are used for the corresponding transliterated letters Ч, Ш, Ю, Я, e.g. “AYuK” (not “AYK”) for Ajvaseda, Yurij Ky`levich (Айваседа, Юрий Кылевич).

Appendix A3 contains the full list of Nenets speakers who contributed to the corpus along with their codes.

2.5.5. Abbreviations used in metadata

AAV: Arkhipov, Alexandre

BJ: Budzisch, Josefina

DCh: Däbritz, Chris Lasse

KAA: Kozlov, Aleksey

KiN: Kim, Natalia

LE: Lazarenko, Elena

PCh: Petschallies, Christiana

SaS: Saldyr, Symbat

WA: Wamprechtshammer, Anna

WNB: Wagner-Nagy, Beáta

2.5.6. Transliteration of Cyrillic names

In the metadata fields referring to personal names and placenames (see 2.8), a romanized spelling is used alongside their Cyrillic spelling according to the Russian orthography.

The following placenames use (a variant of) traditional English spelling:

Arkhangelsk
Leningrad (historical name of Saint Petersburg)
Saint Petersburg
Salekhard
Taimyr

All personal names and most placenames in the metadata are transliterated following the GOST 7.79–2000 System B transliteration standard (GOST 2001; see Appendix A2, Table 28).

Elsewhere, e.g. in text titles, English glosses and free translation, English-style romanization is used (see Appendix A2, Table 29).

2.6. Technical formats

2.6.1. Transcripts

The transcripts in the corpus are provided in several formats, all of them in XML. The main working format is EXMARaLDA EXB, while the other formats are derived from EXB to provide search functionalities and alternative ways of access to the data.

EXMARaLDA EXB and EXS

The annotated transcripts are delivered in the formats of the EXMARaLDA software suite. 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".

Please refer to EXMARaLDA documentation for introduction to the use of this software:
https://exmaralda.org/en/quickstart-documents/.

Time alignment (synchronization)

The transcripts in the corpus are time-aligned with the available sound recordings. Please be aware that the time alignment is only valid at sentence level (**ts** tier; see 3.2.1). Technically, time values are also present at word level (**tx** tier), however they should be disregarded as arbitrary. Time values are also technically present in transcripts without any available sound; these are completely arbitrary and should likewise be disregarded.

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. The fields used are explained in 2.8.

2.6.3. Media

For the text with audio source (AAK_200311_MyLife_nar), the sound recording is provided in Linear PCM WAV format (file extension ".wav") as 22.05 kHz 16 bit mono file ("standard" package).

In the "lite" package, the MP3 version of the sound file is provided (44.1kHz, 192kbps).

2.6.4. Other data

For most texts, source(s) of their original transcription are available as PDF files.

2.7. Workflow of the source files

2.7.1. Transcripts

While the main part of analysis of all texts was done in the *SIL Fieldworks Language Explorer* (FLEx)[6] and all the transcripts were exported to the EXB format via the FLEx's Flextext XML format, previous workflow was different depending on the source of the text. The main steps were digitization, sentence-by-sentence alignment of transcription and translation(s), and the conversion to the project's transcription system (see 3.2.3), followed by import into FLEx.

The format of the source texts from various origins, as well as the method of digitization, is presented in the following table. The digitization process varied depending on the format of the source materials. Some texts were provided already in digital form as Word documents, others were obtained by extracting text from the available PDFs. Some others required OCR processing or were typed manually. The table also indicates whether the original transcript is in Latin script (**stl**) or Cyrillic script (**st**). If only the Latin script version is available, the **st** tier remains empty; otherwise, the Cyrillic version has been transliterated into Latin script (see also section 3.2.2). Additionally, the original Russian, German, or Hungarian translation lines (**ltr**, **ltg**, **lth**) were copied directly from the original files in their available form.

Table 3. Source formats and digitization methods

Source	Source format / digitization method	Latin (stl) or Cyrillic (st) script
Barmich 2018	Typed	st
Burkova 2008	MS Word	st

[6] https://software.sil.org/fieldworks/, last access: 12.06.2024.

Source	Source format / digitization method	Latin (stl) or Cyrillic (st) script
Burkova 2012	OCR	st
Burkova et al. 2003	Toolbox	st
Hajdú 1968	Typed	stl
Koshkareva et al. 2007	OCR	st
Labanauskas 2001	PDF > text	st
Logany & Logany 2016	PDF > text	st
Lyublinskaya 2022	PDF > text	st
Pusztay 1976	OCR	stl
Tereshchenko 1956	OCR	st
Tereshchenko 1990	OCR	st
Turutina 2003	PDF > text	st
Yangasova 2018	OCR	st
Burkova (unpublished)	MS Word	st

2.7.2. Media

The audio file available in the corpus was provided to the project in the WAV format it is published here in.

2.7.3. Metadata

The communication and speaker metadata were extracted from the source publications and information provided by other researchers.

2.8. Metadata for the corpus

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

Personal names and place names are given in their romanized spelling alongside Cyrillic spelling according to Russian orthography. See 2.5.6 on the use of transliteration in these fields.

2.8.1. General corpus metadata

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

2.8.2. Text ("communication") metadata

Name: The code which is assigned to the text (see 2.5.3)

Description:

- **0a. Title:** Short title (in English)
- **0b. Title (RU):** Short title (in Russian)
- **1. Genre:** Abbreviation of the genre of the communication (el= elicitation, flk = folklore, nar = narrative, song = song)
- **2a. Recorded by**: Abbreviation or Name of the person by whom the communication was recorded (see 2.5.5)
- **2b. Date of recording:** Here the date of recording is given (year only)
- **3a-c. Variety / Dialect group / Dialect:** If known, information on the dialect used by the speaker(s) is given here (see 1.2.3); if not, the field remains empty; for Forest Nenets, Dialect Group is not relevant

- **4. Speakers:** Code of the speaker(s)
- **5a. Transcribed by:** Code of the person who did the transcription
- **5b. Date of transcribing:** The exact date (if known) of the transcribing (for written materials, it is the same as the date of collection; for audio materials, the date of the transcribing fieldwork session)
- **5c. Typed by:** Code of the person who did the typing from the manuscript, PDF to text conversion or OCR
- **5d. Time-aligned by:** Code of the person who time-aligned the transcription (if done separately from the transcription)
- **6a-e. Translation into Russian (orginial and edited) / English / German / Hungarian:** Code of the person who did the translation in question.
- **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 2.5.5)
- **8. Glossed by:** Code of the person who did the morphological glossing

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

- **Country:** All the texts were recorded in Russia.
- **Region:** The Region of the recording
- **Settlement:** The place of the recording
- **Settlement (RU):** The place of the recording in Russian
- **Settlement (LatLng):** Geographic coordinates (latitude, longitude) of the settlement

Languages:

- **Language code:** The ISO-code of the language of communication (always *yrk* – Nenets).

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

- **Archive volume:** If the text is taken from an archive, it is mentioned here
- **Published in:** If the text has been published, the publication reference is provided here
- **Published in (bibtex):** If the text has been published, the BiBTeX key of the corresponding entry in the INEL Bibliography is given here

Recording: If an audio file is available, it is linked to the communication description

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 the original archive pages, scans of text publications), they are linked to the communication description here.

2.8.3. Speaker metadata

Metadata of speaker(s) include, on the one hand, biographical information of the speaker and, on the other hand, information on their sociolinguistic background. The level of detail is determined by the available information. Name fields exist both in Russian (RU) and English (translated or transliterated) version. The following fields are defined:

Sigle: Speaker code as defined in 2.5.4

Pseudo: Name shown in Coma's main view (using family name, first name and patronymic)

Sex: male or female

Description:

- **1a-b. Family name (EN, RU)**
- **2a-b. Given name (EN, RU)**
- **3a-b. Patronymic (EN, RU)**
- **4. Also known as:** Unofficial names are given here. These can be traditional names following pre-Christian tradition or nicknames.

Basic biographic data: Here basic biographical data of the speaker are provided. Note that settlement names can refer not always strictly to the settlement itself, but also to tundra camps around this settlement.

- **1a-b. Place of birth (EN, RU)**
- **1c. Place of birth (LatLng)**: Geographic coordinates (latitude, longitude) of place of birth
- **2. Region**
- **3. Country**: (always Russia)
- **4. Date of birth**
- **5. Date of death**
- **6a-b. Grown up in / former residences (EN, RU):** If former residences prior to the work with the linguist are known and differ significantly from place of birth, they are mentioned here
- **7a-b. Domicile (EN, RU)**: The current (i.e. at the time of the recording) place of residence of the speaker if known
- **7c. Domicile (LatLng)**: Geographic coordinates (latitude, longitude) of domicile
- **8a-b. Other information (EN, RU)**

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 if known.
- **2. Higher education:** If the speaker had a higher education, it is sometimes mentioned here.
- **3. Occupation:** Here the profession and/or occupation of the speaker is sometimes mentioned if known.

Family: Here information about the ethnicity of the speaker (in the majority of cases, Nenets) and their family members is given if known. For women, their maiden names are given in parentheses if known. If a speaker's relative is also a speaker with metadata on their own, their code is also given in parentheses.

- **1. Ethnicity**
- **2a-b. Ethnicity of mother / Name of mother**
- **3a-b. Ethnicity of father / Name of father**
- **4a-b. Ethnicity of husband/wife / Name of husband/wife**
- **5a-b. Ethnicity of grandparents / Name of grandparents**
- **6a-b. Family (EN, RU):** Name of the Nenets traditional clan the speaker belongs to is given here if known

Language documentation activities

- **Consultant of:** Here the linguists or other researchers with whom the speaker worked are mentioned. If known, this information concerns also work beyond the texts included in the corpus.

Languages: Here the language repertoire of the speaker is given; Russian (most commonly L2) is not mentioned.

L1 (First language)

- **Language code:** Here the ISO code is given (*yrk* – Nenets).
- **1-4. First language / Variety / Dialect group / Dialect:** Dialectal attribution of the speaker according to the classification in 1.2.3.

L2 (Second language)

- **Language code:** Here the ISO code is given (*kca* – Khanty, *kvp* - Komi).
- **Second language**

3. Transcription and annotation

Many ideas and principles of transcription and annotation go back to the Nganasan Spoken Language Corpus (NSLC) [Brykina et al. 2018], a documentation of this are the respective user guidelines [Wagner-Nagy et al. 2018]. This holds especially true for the annotation principles and annotation schemes for the annotation of semantic roles (SeR), syntactic functions (SyF) and information status (IST), as will be shown in the respective sections.

3.1. Tier layout

Table 4. Tiers in EXMARaLDA files of INEL Nenets Corpus

Tier label	Tier full name	Description	Unit	Optionality
ref	Reference	Text ID + sentence number Text ID + speaker code + sentence number (for texts with multiple speakers)	sentence	obligatory
st	Source transcription (Cyrillic)	Original transcription / orthography of the source if written in Cyrillic alphabet	sentence	obligatory
stl	Source transcription (Latin)	Original transcription of the source if written in Latin alphabet, or original Cyrillic transcription / orthography converted into Latin	sentence	obligatory
ts	Text (sentence)	Main transcription, adapted from stl tier for glossing and annotation.	sentence	obligatory
tx	Text (word)	Main transcription segmented by word	word	obligatory
mb	Morpheme breaks	Morpheme breakdown of words (hyphen-separated morphemes for each word)	morph	obligatory
mp	Morphemes (lexical)	Lexical representation of morphemes	morph	obligatory
ge	Gloss (English)	Morphological glosses (with lexical glosses in English)	morph	obligatory
gr	Gloss (Russian)	Morphological 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
geo	Geographical coordinates	Coordinates of named places and geographical objects	word	optional
SeR	Semantic role	Semantic (thematic) roles of NPs	word / group of words	optional
SyF	Syntactic function	Syntactic functions of predicates and arguments, as well as of subordinate clauses	word / group of words	optional
IST	Information status	Information status of NPs (given/new/accessible)	word	optional
BOR	Borrowing	Borrowings (source language and borrowing type)	word	optional
BOR-Phon	Borrowing phonology	Phonological adaptations in borrowings	word	optional

Tier label	Tier full name	Description	Unit	Optionality
BOR-Morph	Borrowing morphology	Morphological adaptations in borrowings	word	optional
CS	Code switching	Code switching and calques (source language and type)	group of words	optional
ExLocPoss	Existential, locative and possessive structures	Existential, locative and possessive structures	sentence	optional
fr	Free translation (Russian)	Free translation (Russian)	sentence	obligatory
fe	Free translation (English)	Free translation (English)	sentence	obligatory
fg	Free translation (German)	Free translation (German)	sentence	obligatory
fh	Free translation (Hungarian)	Free translation (Hungarian)	sentence	optional
ltr	Original translation (Russian)	Original Russian translation, as provided in the source.	sentence	optional
ltg	Original translation (German)	Original German translation, as provided in the source.	sentence	optional
lth	Original translation (Hungarian)	Original Hungarian translation, as provided in the source.	sentence	optional
nt	Notes	Notes from corpus developers (in English)	sentence	optional
nto	Notes (original)	Original notes from the source	sentence	optional

3.2. Transcription tiers

3.2.1. Main transcription tiers (tx and ts)

The tiers **tx** and **ts** are the main transcription tiers. They use the INEL transcription (see 3.2.3). The major difference between them is that **ts** presents transcriptions of entire sentences, while **tx** has the same content divided into words. Technically speaking, in EXMARaLDA format it is only the **tx** tier which has the type "transcription", all other tiers being of the type "annotation". It is thus the **tx** tier which serves as the basis for segmentation (in "segmented transcription" format, EXS), which is relevant for search using the EXAKT tool and for all sentence and word counts.

The treatment of some special cases and phenomena such as uncertainties and alternatives in transcription, largely follow general INEL conventions described in (Arkhipov 2020: 12–17). Some kinds of uncertainties are marked informally in notes tiers (**nt** and/or **nto**).

(1)

ref	AE_1957_ThreeFriends_flk.001 (001)		
ts	n'aaxarʔ n'aa ŋaewii.		
tx	n'aaxarʔ	n'aa	ŋaewii.

3.2.2. Source transcription tiers (st and stl)

For the texts coming from a source using Cyrillic transcription / orthography, the source transcription tier (**st**) contains the original Cyrillic version of the text (otherwise it is empty); in this case the **stl** tier may contain this original phonetic transcription converted into Latin alphabet. For the texts coming from a source using Latin transcription, the **st** tier is always empty, and the **stl** tier contains the original transcription.

(2)

ref	ANK_19980705_Orphan_flk.004 (001.004)
stl	Ŋajɬ'a n'anantuŋ t'iɬ'iŋa.
st	Ӈайӆя нянантуӈ тиӆиӈа.

3.2.3. Transcription and orthographical conventions in the corpus

For transliteration of Cyrillic names in the metadata, see 2.5.6.

INEL transcription system

The transcription used in the corpus generally represents a rendering of the original transcriptions / orthographical representation in the sources into a unified set of Latin transcription symbols. It thus differs significantly from Salminen's deep (morpho)phonological transcription (e.g. as represented in Salminen 1998).

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

- *ŋ* is used for the voiced velar nasal
- *ɬ* is used for the voiceless alveolar lateral fricative
- *t', d', n', l', ɬ'* are used for palatalized or palatal consonants
- *č* is used for the postalveolar/palatalized affricate corresponding to orthographic <ч>
- *š* is used for the postalveolar/palatalized fricative corresponding to orthographic <ш>
- *ä* is used for the /ae/ diphthong in NenF corresponding e.g. to orthographic <ӭ>
- *ʔ* is used for both (morpho)phonologically distinct glottal stops
- double characters are used for long vowels, e.g. *oo* for /oː/

For particular correspondences between original transcriptions used in the sources and the INEL transcription see Appendix A1.

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 text name and the number of the sentence, separated by a full stop. The sentences are numbered throughout the entire text. The sentence numbers are zero-padded up to 3 digits. This part of the **ref** tier should be used for citation of a specific sentence coming from the corpus.

In brackets, the numbering according to the FLEx scheme is given (*paragraph_number*.*sentence_number*). The FLEx number is only kept for internal error tracking and has neither linguistic value nor relevance for citation. Thus e.g. the sentences in the example below should be cited as "YVD_1976_OldKhantysSon_flk.001" and "YVD_1976_OldKhantysSon_flk.002" respectively:

(3)

ref	YaVD_1976_OldKhantysSon_flk.001 (001.001)	YaVD_1976_OldKhantysSon_flk.001 (001.002)

In texts recorded from multiple speakers, the speaker code is additionally provided between the text code and the sentence number, separated by dots. The numbering is consecutive within each speaker separately, starting from 001. (The Flex numbering is consecutive across speakers.)

(4)

ref-VA	VA_VR_200206_Wandering_nar.VA.018 (018)		VA_VR_200206_Wandering_nar.VA.019 (020)
ref-VR		VA_VR_200206_Wandering_nar.VR.001 (019)	

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 and the derived stem is then glossed as a separate lexical item.

(5)

ref	LACh_XX_HalfFace_flk.037 (001.037)		
ts	(Katɬ’uʔ tesamiʔ ŋ’iʔn’išaʔ).		
tx	(Katɬ’uʔ	tesamiʔ	ŋ’iʔn’išaʔ).
mb	katɬ’uʔ	te-sami-ʔ	ŋ’i-ʔn’išaʔ
fe	(They must have had many reindeer.)		

3.3.3. Morphemes (lexical) (mp)

The underlying morphemes tier (**mp**) shows the lexical representation of the morphs, both stems and affixes, which appear separated in the **mb** tier. Stems are, hence, represented here by their lexical entry in the FLEx lexicon. Affixes are represented by their main allomorphs. All morphemes within a word are separated by hyphens. Zero morphs are not represented in this tier.

(6)

ref	LACh_XX_HalfFace_flk.037 (001.037)		
ts	(Katɬ’uʔ tesamiʔ ŋ’iʔn’išaʔ).		
tx	(Katɬ’uʔ	tesamiʔ	ŋ’iʔn’išaʔ).
mb	katɬ’uʔ	te-sami-ʔ	ŋ’i-ʔn’išaʔ
mp	katɬ’uʔ	ti-sami-ʔ	ŋi-ʔn’iša
fe	(They must have had many reindeer.)		

3.3.4. Gloss (ge, 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.[7] For the full list of glossing abbreviations, see Appendix A4.

Glosses for all morphemes within a word are separated with hyphens. Glosses for zero morphs are given in square brackets preceded by a dot (e.g. ".[NOM.SG]").

If a gloss contains two or more semantic components, these are separated by a dot. For more convenient reading the dot is omitted in combinations of person and number (e.g. "IMP.2SG").

Morphemes with unknown meaning are glossed with two percent signs (%%). One leading percent sign indicates that the gloss is tentative (e.g. "%whistle").

[7] https://www.eva.mpg.de/lingua/resources/glossing-rules.php, last access: 22.12.2023.

(7)

ref	LACh_XX_HalfFace_flk.037 (001.037)		
ts	(Katɬ'uʔ tesamiʔ ŋ'iʔn'išaʔ).		
tx	(Katɬ'uʔ	tesamiʔ	ŋ'iʔn'išaʔ).
mb	katɬ'uʔ	te-sami-ʔ	ŋ'i-ʔn'išaʔ
mp	katɬ'uʔ	ti-sami-ʔ	ŋi-ʔn'iša
ge	probably	deer-PROP-NOM.PL	be-PROB.[3SG.S]
gr	наверное	олень-PROP-NOM.PL	быть-PROB.[3SG.S]
fe	(They must have had many reindeer.)		

3.3.5. Morphological category (mc)

The **mc** tier indicates the morphological category of both lexical stems (i.e. the part of speech) and affixes (i.e. the inflectional category or the derivational process). For inflectional affixes the pattern "x:a" is used, where x stands for part of speech, to which an affix can be attached, and a stands for the category of this affix. Derivational processes are marked as "x>y", x and y being the tags for part of speech. Elements with unknown meaning are marked with two percent signs ("%%"). For the list of abbreviations, see Appendix A5.

(8)

ref	LACh_XX_HalfFace_flk.037 (001.037)		
ts	(Katɬ'uʔ tesamiʔ ŋ'iʔn'išaʔ).		
tx	(Katɬ'uʔ	tesamiʔ	ŋ'iʔn'išaʔ).
mb	katɬ'uʔ	te-sami-ʔ	ŋ'i-ʔn'išaʔ
mp	katɬ'uʔ	ti-sami-ʔ	ŋi-ʔn'iša
ge	probably	reindeer-PROP-NOM.PL	be-PROB.[3SG.S]
gr	наверное	олень-PROP-NOM.PL	быть-PROB.[3SG.S]
mc	adv	n-n>adj-n:(num)	v-v:(mood).[v:pn]
fe	(They must have had many reindeer.)		

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 the same as in the morphological category tier **mc**.

(9)

ref	LACh_XX_HalfFace_flk.037 (001.037)		
ts	(Katɬ'uʔ tesamiʔ ŋ'iʔn'išaʔ).		
tx	(Katɬ'uʔ	tesamiʔ	ŋ'iʔn'išaʔ).
mb	katɬ'uʔ	te-sami-ʔ	ŋ'i-ʔn'išaʔ
mp	katɬ'uʔ	ti-sami-ʔ	ŋi-ʔn'iša
ge	probably	deer-PROP-NOM.PL	be-PROB.[3SG.S]
gr	наверное	олень-PROP-NOM.PL	быть-PROB.[3SG.S]
mc	adv	n-n>adj-n:(num)	v-v:(mood).[v:pn]
ps	adv	n	v
fe	(They must have had many reindeer.)		

3.3.7. Geographical coordinates (geo)

For some placenames encountered in the corpus, geographical coordinates in "latitude, longitude" format are provided in the **geo** tier.

(10)

ref	RE_197012_Childhood1_nar.001 (001.001)			
ts	man'a kaɬ'tama num too.			
tx	man'a	kaɬ'tama	num	too.
ge	PRO1PL.[NOM.SG]	village-NOM.SG.1PL	god.[NOM.SG]	lake.[NOM.SG]
geo			63.513633, 71.357055	
fe	Our village is called Num To.			

3.3.8. Syntactic function (SyF)

The annotation scheme used in the syntactic function tier was developed by Beáta Wagner-Nagy and Sándor Szeverényi [Wagner-Nagy et al. 2018: 21ff.] who also made it available for the project.

In the Syntactic function tier (SyF) basic syntactic functions (i.e. subject, direct object, predicate) are tagged. The form of the tag is similar to the tags used in the Semantic Roles tier (SeR): <form.animacy:syntactic function>. Subjects and direct objects are tagged at the head of the respective phrase, zero subjects and objects are tagged at the predicate of the clause. For complex verbal predicates the cells of the main verb and the auxiliary are merged. The following tags are used:

Table 5. Tags for annotating syntactic functions

Abbreviation	**Description**
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
0.3.h:O	zero/covert third-person human object
0.3:O	zero/covert third-person non-human object
Predicate	
v:pred	verbal predicate
n:pred	nominal predicate
adj:pred	attributive/adjectival predicate
pro:pred	pronominal predicate
ptcl:pred	particle predicate
cop	copula

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 6. Tags for annotating subordinate clauses

Abbreviation	**Description**
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.*)

Here are some examples for tagging syntactic functions:

(11)

ref	RE_197012_Childhood_nar.001 (001.001)			
ts	man'a kaɬ'tama num too.			
tx	man'a	kaɬ'tama	num	too.
mb	man'a	kaɬ'ta-ma	num	too
ge	PRO1PL.[NOM.SG]	village-NOM.SG.1PL	god.[NOM.SG]	lake.[NOM.SG]
SyF		np:S		
fe	Our village is called Num To.			

(12)

ref	LACh_XX_HalfFace_flk.037 (001.037)		
ts	(Katɬ'uʔ tesamiʔ ŋ'iʔn'išaʔ).		
tx	(Katɬ'uʔ	tesamiʔ	ŋ'iʔn'išaʔ).
mb	katɬ'uʔ	te-sami-ʔ	ŋ'i-ʔn'išaʔ
ge	probably	deer-PROP-NOM.PL	be-PROB.[3SG.S]
SyF		np:S	v:pred
fe	(They must have had many reindeer.)		

(13)

ref	YaVD_1981_SonOfGod_flk.012 (001.012)		
ts	S'iw xom maneŋa.		
tx	S'iw	xom	maneŋa.
mb	s'iw	xo-m	mane-ŋa
ge	seven.[NOM.SG]	birch-ACC.SG	see-CO.[3SG.S]
SyF		np:O	0.3.h:S v:pred
fe	He saw seven birches.		

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 [Haig & Schnell 2014] with further developments by Beáta Wagner-Nagy and Sándor Szeverényi [Wagner-Nagy et al. 2018: 21ff.], adapted for the current 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 whole phrase, then the semantic role is tagged at the head of the phrase. In postpositional constructions the cells of the postposition and its complement are combined. Zero referents are tagged by default at the predicate of the sentence. Semantic roles are tagged both in main and in dependent clauses. In the "animacy" category, human and non-human referents are differentiated. Human referents additionally get the tag <.h>, non-human referents get no marking in this category. The following table summarizes the tags used for referent expressions:

Table 7. Tags for referent expressions

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

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

Table 8. Semantic Roles tagged and their abbreviations

Semantic Role	**Abbreviation**	**Description**
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. 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 - arguments of unaccusative verbs such as *die, fall*
Possessor	Poss	entity which owns something both alienable and in-alienable possession - also inanimate referents (e.g. the top of the mountain)
Recipient	R	(mostly animate) recipient of transfer of something - addressee of verba dicendi
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 actant of verbs like *see*, *hear*, *feel*, but NOT of verbs like *look for*, *listen*
Theme	Theme	entity which is moved or affected by some action (change of location or possession; object of transfer) entity whose location is specified - object of possession (possessee)
Time	Time	- time point or an interval of time

The following charts show some examples of tagging Semantic Roles:

(14)

ref	YaVD_1981_SonOfGod_flk.012 (001.012)		
ts	S'iw xom maneŋa.		
tx	S'iw	xom	maneŋa.
mb	s'iw	xo-m	mane-ŋa
ge	seven.[NOM.SG]	birch-ACC.SG	see-CO.[3SG.S]
SeR		np:Th	0.3.h:E
fe	He saw seven birches.		

(15)

ref	AE_1957_ThreeFriends_flk.025 (025)			
ts	warŋer maa: man’ ŋawortaaw.			
tx	warŋer	maa:	man’	ŋawortaaw.
mb	warŋe-r	maa	man’	ŋaw-o-r-ta-a-w
ge	crow-NOM.SG.2SG	say.[3SG.S]	PRO1SG.[NOM]	eat-EP-FRQ-IPFV-CO-1SG.SG.O
SeR	np.h:A		pro.h:A	
fe	The crow said: "I will eat it."			

(16)

ref	TPG_2002_Wella_flk.059 (001.059)		
ts	Kukaxäna m’aʔkna me.		
tx	kukaxäna	m’aʔkna	me.
mb	kukaxäna	m’aʔ-kna	me
ge	once	tent-LOC.SG	be.there.[3SG.S]
SeR		np:L	0.3.h:A
fe	Once she remained in the tent.		

3.3.10.Information status (IST)

The Information status tier (IST) contains the annotation of information status. The annotation is based on the annotation guidelines for information structure and information status in [Götze et al. 2007], some minor changes were nevertheless done. The principles of annotation and the annotation scheme itself were developed by Wagner-Nagy & Szeverényi [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 9. Parameters for determining information status

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

In detail that 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 10. Basic tags for annotating information status

Tag	**Description**
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

As Nenets is a pro-drop language, many referents are not overtly realized in the sentence. Therefore the information status also of non-overt referents is tagged. The tag set remains the same, the prefix <0.> being added to the tag in question (e.g. <0.giv-active> for a zero/covert given and active referent), and the referent is tagged at the predicate of the clause.

(17)

ref	YaVD_1981_SonOfGod_flk.012 (001.012)		
ts	S'iw xom maneŋa.		
tx	S'iw	xom	maneŋa.
mb	s'iw	xo-m	mane-ŋa
ge	seven.[NOM.SG]	birch-ACC.SG	see-CO.[3SG.S]
IST		new	0.giv-active
fe	He saw seven birches.		

(18)

ref	AE_1957_ThreeFriends_flk.025 (025)			
ts	warŋer maa: man' ŋawortaaw.			
tx	warŋer	maa:	man'	ŋawortaaw.
mb	warŋe-r	maa	man'	ŋaw-o-r-ta-a-w
ge	crow-NOM.SG.2SG	say.[3SG.S]	PRO1SG.[NOM]	eat-EP-FRQ-IPFV-CO-1SG.SG.O
IST	giv-active		giv-active	
fe	The crow said: "I will eat it."			

(19)

ref	TPG_2002_Wella_flk.059 (001.059)		
ts	Kukaxäna m'aʔkna me.		
tx	kukaxäna	m'aʔkna	me.
mb	kukaxäna	m'aʔ-kna	me
ge	once	tent-LOC.SG	be.there.[3SG.S]
IST		accs-sit	0.giv-active
fe	Once she remained in the tent.		

3.3.11.Borrowings and code switching (BOR, BOR-Phon, BOR-Morph, CS)

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 Nenets, Russian (RUS) borrowing is annotated. For the type of borrowing the following tags are used (cf. also Arkhipov 2020):

Table 11. Tags for annotating borrowings

Tag	**Description**
: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

In several cases, a loan translation is signaled by the "RUS:calq" tag in the BOR tier.

3.3.11.1. Borrowing phonology (BOR-Phon)

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	Description
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)

3.3.11.2. Borrowing morphology (BOR-morph)

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	Description
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. an inflected paradigm item is borrowed)
Further inflection (in the matrix language)	
:bare	no inflection
:infl	further inflection

An example of annotations in BOR-morph tier follows:

(20)

ref	YaNN_1976_SonOfGod_flk.023 (001.023)			
ts	Xuw m’erčawna jedamt p’ir’ejin.			
tx	Xuw	m’erčawna	jedamt	p’ir’ejin.
mb	xuw	m’erčawna	Jeda-m-t	p’ir’e-ji-n
ge	in.the.morning	early	food-ACC.SG-OBL.2SG	cook-OPT-2SG.S
BOR			RUS:core	
BOR-morph			dir:infl	
fe	Cook some food early in the morning!			

3.3.11.3. Code-switching (CS)

The Code switching tier (CS) contains the annotation of code-switching. Whereas borrowings treat single words, code switching (mostly) treats sequences of two or more words. Both language of the code-switch and type of the code switch are annotated, namely according to the scheme <LANGUAGE:type>. The language is mostly Russian (RUS). The tag set for the type of code-switch is the following:

Table 14. Tags for annotating code-switching

Tag	Description
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. an NP or a PP is inserted)
:int.alt	the point of change is somewhere at an arbitrary point in the sentence
:int	a single word is inserted, distinguishing between subtypes is problematic

3.3.12. ExLocPoss

The ExLocPoss tier provides the annotation of existential, locative and possessive clauses. Existential and locative clauses express the episodic presence/absence of a referent X (figure) at a place Y (ground). In locative predication, the figure serves as the starting point for the perspectivization of the state of affairs, whereas it is the ground in existential predication. As a corollary, the figure is prototypically definite and topical in locative predication, whereas it is prototypically indefinite in existential predication, belonging to the focus domain. Possessive clauses 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 core structures are thus the translational equivalents of:

"At Y, there is (no) X" (existential), e.g. *There is a book on the table.*

"X is (not) at Y" (locative), e.g. *The book is on the table.*

"X has (no) Y" (possessive), e.g. *The girl has a book.*

Additionally, generic existential clauses (see Koch 2012; a.k.a. *hyparctics* in Haspelmath 2022) can predicate the permanent presence or existence of the referent X (figure), e.g. *there is a/no God* or *there is beer without alcohol.* These are subsumed under existential clauses in the annotation. Inverted possessive clauses, i.e. appertentive clauses in Haspelmath's (2022) terms (e.g. English *the book is John's*), are subsumed under possessive clauses in the annotation.

The annotation scheme includes the three functional domains existential (Ex), locative (Loc) and possessive (Poss), the coding strategy (see below), as well as the polarity (Aff or Neg) of the clause. The annotation tags have the format **Domain:Strategy.Polarity**. Table 12 lists the tags used for annotating existential, locative and possessive clauses.

Table 15. Tags for annotating existential, locative and possessive clauses

Tag	Comment
Functional domain	
Ex	existential clause
Loc	locative clause
Poss	possessive clause
Coding strategy and polarity	
:Zero.Aff	no lexical linking element, does not exclude pn-marking; affirmative
:Zero.Neg	no lexical linking element, does not exclude pn-marking; 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

As shown in Table 14, "zero copula" includes instances with and without pn-agreement at the predicate, i.e. the ground or possessee element. "Copula" means any linking item that has bleached semantics, regardless of whether it is nominal or verbal; most often, this item occurs in nominal or adjectival predication, as well. For tagging a linking element as "existential item", it must be clear from a language-internal perspective that the item indeed has

existential semantics. Posture verbs are translational equivalents of English *sit*, *stand*, *lie*, *hang* etc.; note that clauses including posture verbs are only annotated as existential or locative clauses, if the figure element is inanimate. This is due to the fact that animate figures most often imply a semantic description of the posture instead of merely predicating episodic presence/absence (see Däbritz 2023 for a discussion). *have*-verbs are transitive translational equivalents for English *have*, invenitive verbs are verbs formed from a root meaning 'find', and other verbs subsume mostly verbs like 'be situated' or alike; these three possibilities are not encountered in the corpus.

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

(21)

ref	YaVD_1976_YoungShaman_flk.137 (001.137)		
ts	Jir'ir p'ix'in'a me.		
tx	Jir'ir	p'ix'in'a	me.
mb	jir'i-r	p'i-x'in'a	me
ge	grandfather-NOM.SG.2SG	outside-LOC.PL	be.there.[3SG.S]
ExLocPoss	Loc:Cop.Aff		
fe	Your grandfather is outside.		

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

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

(22)

ref	YaVD_1981_SonOfGod_flk.012 (001.012)		
ts	S'iw xom maneŋa.		
tx	S'iw	xom	maneŋa.
mb	s'iw	xo-m	mane-ŋa
mp	s'iʔiw	xo-mʔ	manes-ŋa
ge	seven.[NOM.SG]	birch-ACC.SG	see-CO.[3SG.S]
gr	семь.[NOM.SG]	береза-ACC.SG	увидеть-CO.[3SG.S]
fe	He saw seven birches.		
fg	Er sah sieben Birken.		
fr	Он увидел семь берёз.		

3.3.14.Literal translation (ltr, ltg, lth)

Texts published with a Russian, German or Hungarian version have this translation added; it can be found in the respective tier.

3.3.15.Notes (nt, nto)

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 2.5.5, in square brackets, followed by a colon).

3.4. Searching the corpus

3.4.1. Search with EXMARaLDA EXAKT

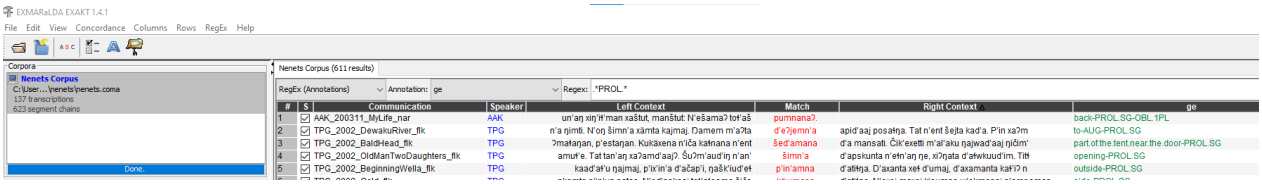
The EXMARaLDA software suite includes EXAKT, an analysis and concordance tool.

In order to perform a search on the downloaded corpus files locally, the main metadata file (**nenets.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[8]) 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 A4 and Appendix A5 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[9] 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/NenetsCorpus/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)
- 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 “Nenets” and the search expression(s) entered in one or more corresponding fields.

Table 16. Tsakorpus search fields and EXMARaLDA tiers: main transcription and word-/morph-level annotation

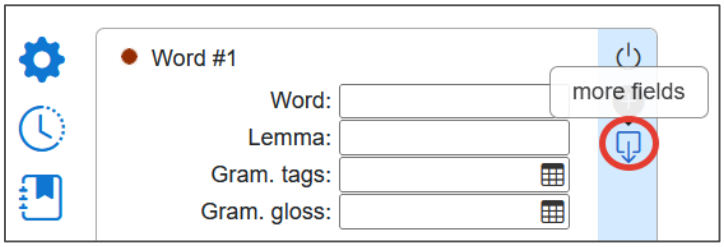
Tsakorpus search field	**Corresponding tier in EXMARaLDA**
Word	**tx**
Lemma	**mp** (stem)

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

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**
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



Please refer to sections 3.2–3.3, Appendix A4 and Appendix A5 for annotations used in the corpus.

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. "ti") in the **Lemma** field, or by searching for its lexical gloss (e.g. "reindeer" / "олень") in **Lex. gloss (en)** or **Lex. gloss (ru)** fields.

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

To find only a particular allomorph, its form can be specified in curly braces following the gloss in the **Gram. gloss** field: "ABL.PL{xat}".

In case there exist more than one underlying form of an affix in **mp** tier (e.g. two dialectal variants), a particular underlying form can be specified in curly braces with underscore in the **Gram. gloss** field: "ABL.PL{_kăt}".

Table 17. Stems and affixes in Tsakorpus

EXMARaLDA tier	**Word**	**Stem**	**Search field**	**Affix**	**Search field**
tx	tixat				
mb	ti-xat	ti		xat	Gram. gloss: ABL.PL{xat}
mp	ti-kăt	ti	Lemma: ti	kăt	Gram. gloss: ABL.PL{_kăt }
ge	reindeer-ABL.PL	reindeer	Lex. gloss (en): reindeer	ABL.PL	Gram. gloss: ABL.PL Gram. tags: abl,pl
gr	олень-ABL.PL	олень	Lex. gloss (ru): олень	ABL.PL	Gram. gloss: ABL.PL Gram. tags: abl,pl
ps	n		Gram. tags: n		

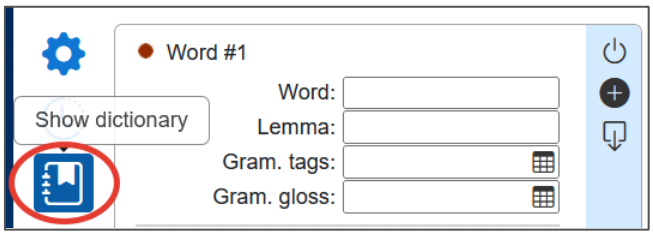
Note that some stems may have a glossing label similar to a grammatical gloss, e.g. NEG for "negative verb". In such cases, this gloss should be entered into **Lex. gloss (en)** or **Lex. gloss (ru)** field, since it will not be treated as a grammatical gloss by Tsakorpus. It will however get a grammar tag (see next section) and can be found by searching for this tag. Such glosses are marked as "lexical" in Comment columns in Appendix A5.

Table 18. Stems and affixes in Tsakorpus: stems with grammatical glosses

EXMARaLDA tier	Word	Stem	Search field	Affix	Search field
tx	n'idam				
mb	n'i-dam	n'i		dam	**Gram. gloss:** 1SG.S{dam}
mp	n'i-dm?	n'i	**Lemma:** n'i	dm?	**Gram. gloss:** 1SG.S {_dm?}
ge	NEG-1SG.S	NEG	**Lex. gloss (en):** NEG **Gram. tags:** neg,negv	1SG.S	**Gram. gloss:** 1SG.S **Gram. tags:** pn1,pnsg,subjc
gr	NEG-1SG.S	NEG	**Lex. gloss (ru):** NEG **Gram. tags:** neg,negv	1SG.S	**Gram. gloss:** 1SG.S **Gram. tags:** pn1,pnsg,subjc
ps	aux		**Gram. tags:** aux		

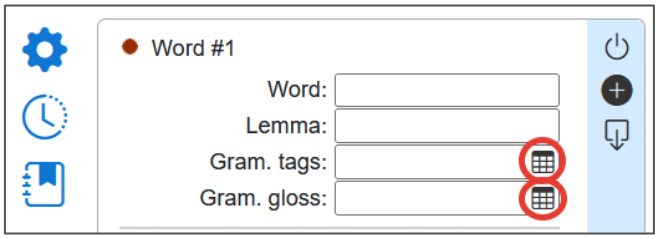
A list of lemmas (i.e. underlying forms of stems as represented in **mp** tier) along with their translations (lexical glosses) can be displayed with "Show dictionary" button.

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 A5.

- 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 A5. 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 A5 for complete lists of tags.

- When a gloss consists of multiple components, such as “ABL.SG” or “3DU.MD”, each of them is usually translated into a tag, e.g. “ABL.SG” => “abl” (ablative), “sg” (singular); “PROL.PL” => “prol” (prolative), “pl” (plural); “3DU.MD” => “pn3” (3 person), “pndu” (dual number), “md” (medial conjugation). A search for tag “pn3” will return all words with any of glosses “3SG”, “3DU”, “3PL”, “3DU.MD”, etc.; a search for tag “abl” will return all words with any of glosses “ABL.SG”, “ABL.PL”.
- When searching with glosses, the entire gloss should be entered as a search expression. E.g. a search for “PROL” will not find “PROL.SG”. Use grammar tags if you need to search for a component of a complex gloss.
- 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. E.g. a search for a gloss “OBL.SG” will not find a zero morph in “house.[OBL.SG]”. Such wordforms can only be found with corresponding grammar tags, e.g. “obl,sg” in this case.
- 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 A4) 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”.

To search in one of the sentence-level annotation tiers, the search expression should be entered into “Word” field, and “Language/tier” field set to one of the following:

Table 19. Tsakorpus search fields and EXMARaLDA tiers: sentence-level annotation

Language/tier label	**Corresponding tier in EXMARaLDA**
Nenets orthography (Cyr.)	**tsc**
Nenets source (Cyr.)	**st**
Nenets source (Lat.)	**stl**
Russian translation	**fr**
English translation	**fe**
Orig. Russian translation	**ltr**
Orig. German translation	**ltg**
Orig. Hungarian translation	**lth**
Notes (Eng.)	**nt**
Orig. notes	**nto**
Sentence ID	**ref**

For further details please refer to Tsakorpus online help.

Figure 6. Tsakorpus interface: Show help

References

Arkhipov, A. V., & Däbritz, C. L. (2018). Hamburg corpora for indigenous Northern Eurasian languages. *Tomsk Journal of Linguistics and Anthropology*, Issue 3 (21), 9–18.

Barmich, M. Ya. [Бармич, М. Я.] (2018). *Словарь языка канинских ненцев: С кратким грамматическим очерком канинского говора ненецкого языка*. Санкт-Петербург: Издательство РГПУ.

Brykina, M., Gusev, V., Szeverényi, S., & Wagner-Nagy, B. (2018). *Nganasan Spoken Language Corpus (NSLC). Archived in Hamburger Zentrum für Sprachkorpora. Version 0.2.* Publication date: 12.06.2018. URL: http://hdl.handle.net/11022/0000-0007-C6F2-8

Burkova, S. I. [Буркова, С. И.] (2010). *Диалектологический словарь ненецкого языка.* Екатеринбург: Баско.

Burkova, S. I. [Буркова, С. И.] (2012). Тексты на надымском говоре тундрового диалекта ненецкого языка. In *Экспедиционные материалы по языкам народов Сибири* (1995–2012 гг.) (pp. 211–260).

Burkova, S. I. [Буркова, С. И.] (2008). Аналитические глагольные формы сказуемого с модальным значением в текстах на лесном диалекте ненецкого языка. *Научный вестник Ямало-Ненецкого АО*, Вып. 6 (58), 104–138.

Burkova, S. I. (2022). Nenets. In M. Bakró-Nagy, J. Laakso, & E. Skribnik (Eds.), *The Oxford Guide to the Uralic Languages* (pp. 674–708). Oxford: Oxford University Press.

Burkova, S. [Буркова, С. И.] (2003). Лесной диалект ненецкого языка (пуровский говор). *Языки и фольклор коренных народов Сибири*, 10, 141–144.

Götze, M., Dipper, S., Götze, M., & Skopeteas, S. (2007). Information structure. In 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: 02.11.2017]

Haig, G., & Schnell, S. (2014). Annotations using GRAID (Grammatical relations and animacy in discourse). *Introduction and guidelines for annotators, Version* 7.0. URL: https://www.uni-bamberg.de/fileadmin/aspra/Publications/GRAID7.0_manual.pdf [Accessed: 01.11.2017]

Hajdú, P. (1968). *Chrestomathia Samoiedica.* Budapest: Tankönyvkiadó.

Koshkareva, N. B. [Кошкарева, Н. Б.] (2005). *Очерки по синтаксису лесного диалекта ненецкого языка: Часть первая. Синтаксические связи.* Новосибирск: Российская академия наук, Сибирское отделение, Институт филологии.

Koshkareva, N. B., Burkova, S. I., & Shilova V. V. [Кошкарева, Н. Б., Буркова, С. И., & Шилова, В. В.] (2007). *Пящитаӈ неша'' вача''мы'': Рассказы варьёганских ненцев. На ненецком (лесной диалект) и русском языках.* Ханты-Мансийск: Изд-во «Полиграфист».

Labanauskas, K. I. [Лабанаускас, К. И.] (2001). *Ямидхы лаханаку - Сказы седой старины.* Русская литература.

Logany, A. Ch., & Logany, A. O. [Логаны, А. Ч., & Логаны, А. О.] (2016). *Сказки Татвы: сборник фольклорных произведений на лесном диалекте ненецкого языка* (Н. Б. Кошкарева, Ed.). Ханты-Мансийск: Издательство ООО «ФОРМАТ».

Lyubinskaya, M. [Люблинская, М.] (2022). Трикстеры-пройдохи в ненецком фольклоре. In V. Gusev, A. Urmanchieva, & A. Anikin (Eds.), Siberica et Uralica: In memoriam Eugen Helimski (pp. 395–414).

Pusztay, J. (1976). Material aus dem Wald-Dialekt des Jurak-Samojedischen (Vol. 10). *Nachrichten der Akademie der Wissenschaften in Göttingen.* Vandenhoeck und Ruprecht.

Salminen, T. (1998). Nenets. In: *The Uralic languages*, ed. by D. Abondolo. London, New York: Routledge. 516–547. (Routledge language family descriptions.)

Tereshchenko, N. M. [Терещенко, Н. М.] (1956). *Материалы и исследования по языку ненцев*. Издательство Академии Наук СССР.

Tereshchenko, N. M. [Терещенко, Н. М.]. (1990). *Ненецкий эпос. Материалы и исследования по самодийским языкам.* Наука.

Turutina, P. G. [Турутина, П. Г.] (2003). *Нешаӈ ванӆат шотпяӆвс'' (Легенды и сказки лесных ненцев).* Новосибирский государственный университет.

Verbov, G. D. [Вербов, Г. Д.] (1973). Диалект лесных ненцев. In *Самодийский сборник* (pp. 4–190). Новосибирск.

Yangasova, N. M. [Янгасова, Н. М.] (2018). *Фольклор ямальских ненцев*.

Appendix A1. Cyrillic to Latin INEL transcription conversion for Nenets

In this Appendix, transliteration rules in TeCKit[10] format are given which were used in conversion from various variants of Cyrillic orthography / transcription into INEL transcription. The resulting transcription in **ts** / **tx** tiers was subsequently manually edited where necessary.[11]

Table 20. Transliteration rules for Burkova 2008

```
pass(Unicode)

UniClass [VowelNonPalC] =     ('ы' 'э' 'а' 'о' 'у') ; 'ы''э''а''о''у'
UniClass [VowelL] =          ('i' 'e' 'a' 'o' 'u') ; 'i''e''a''o''u'
UniClass [VowelNonPalCCap] = ('Ы' 'Э' 'А' 'О' 'У') ; 'Ы''Э''А''О''У'
UniClass [VowelLCap] =         ('I' 'E' 'A' 'O' 'U') ; 'I''E''A''O''U'

; glottal stop

'”' <> latin_letter_glottal_stop ; U+201D (in Burkova-TB)
'’' <> latin_letter_glottal_stop ; U+2019 (in Burkova-TB)

; diphthong

'ӭ' <> 'ä' ; э'□' (U+F50F) in the Word file, looks like 'э '
'Ӭ' <> 'Ä' ; э'□' (U+F50F) in the Word file, looks like 'э '

; plain vowels reverse direction

[VowelNonPalC] < [VowelL]
[VowelNonPalCCap] < [VowelLCap]

pass(Unicode)

; и intervocalically

UniClass [VowelC] = ('ы' 'э' 'а' 'о' 'у' 'и' 'е' 'я' 'ё' 'ю' 'ä' 'Ы' 'Э' 'А' 'О' 'У' 'И' 'Е' 'Я'
'Ё' 'Ю' 'Ä')

[VowelC] cyrillic_small_letter_i <> [VowelC] latin_small_letter_j latin_small_letter_i


pass(Unicode)

; vowels after D, N (palat.) and other vowels

UniClass [ConsPalC] = ('п' 'в' 'м' 'т' 'д' 'н' 'л' 'ԓ' 'к' 'х' 'ӈ' 'ӊ') ; 'ԓ' 'ӊ' (in Burkova-TB)
UniClass [ConsPalCCap] = ('П' 'В' 'М' 'Т' 'Д' 'Н' 'Л' 'Ԓ' 'К' 'Х' 'Ӈ' 'Ӊ') ; 'Ԓ' 'Ӊ' (in Burkova-
TB)
UniClass [ConsPalL] = ('p' 'w' 'm' 't' 'd' 'n' 'l' 'ɬ' 'k' 'x' 'ŋ' 'ŋ') ;
UniClass [ConsPalLCap] = ('P' 'W' 'M' 'T' 'D' 'N' 'L' 'ɬ' 'K' 'X' 'Ŋ' 'Ŋ') ;

UniClass [ConsNonPalC] = ('ч''ш''с') ; ч ш
UniClass [ConsNonPalCCap] = ('Ч''Ш''С') ; Ч Ш
UniClass [ConsNonPalL] = ('č''š''s') ; č š
UniClass [ConsNonPalLCap] = ('Č''Š''S') ; Č Š


UniClass [VowelPalC] =        ('и' 'е' 'я' 'ё' 'ю') ;
UniClass [VowelNonPalC] =     ('ы' 'э' 'а' 'о' 'у') ;
UniClass [VowelL] =          ('i' 'e' 'a' 'o' 'u') ;
UniClass [VowelPalXiC] =      ('e' 'я' 'ё' 'ю') ;
UniClass [VowelXiL] =  ('e' 'a' 'o' 'u') ;

UniClass [VowelPalCCap] =       ('И' 'Е' 'Я' 'Ё' 'Ю') ;
UniClass [VowelNonPalCCap] = ('Ы' 'Э' 'А' 'О' 'У') ;
UniClass [VowelLCap] =          ('I' 'E' 'A' 'O' 'U') ;
UniClass [VowelPalXiCCap] =  ('E' 'Я' 'Ё' 'Ю') ;
UniClass [VowelXiLCap] =        ('E' 'A' 'O' 'U') ;
```

[10] https://software.sil.org/teckit/, last accessed: 17.12.2024.

[11] We are grateful to Tapani Salminen for pointing out a recurrent error in Cyrillic to Latin conversion.

```
[ConsPalC]=c [VowelPalC]=v > [ConsPalL]=c modifier_letter_apostrophe [VowelL]=v
[ConsNonPalC]=c [VowelPalC]=v > [ConsNonPalL]=c [VowelL]=v
[VowelPalXiC]=v <> latin_small_letter_j [VowelXiL]=v
[VowelNonPalC] > [VowelL]
'и' > 'i'

[ConsPalCCap]=c [VowelPalC]=v > [ConsPalLCap]=c modifier_letter_apostrophe [VowelL]=v
[ConsNonPalCCap]=c [VowelPalC]=v > [ConsNonPalLCap]=c [VowelL]=v
[VowelPalXiCCap]=v <> latin_capital_letter_j [VowelXiL]=v
[VowelNonPalCCap] > [VowelLCap]
'И' > 'I'

[VowelPalC]=v < modifier_letter_apostrophe [VowelL]=v
[VowelPalCCap]=v < modifier_letter_apostrophe [VowelLCap]=v


pass(Unicode)

; long vowels

'i' combining_macron <> 'ii'
'e' combining_macron <> 'ee'
'a' combining_macron <> 'aa'
'o' combining_macron <> 'oo'
'u' combining_macron <> 'uu'
'ä' combining_macron <> 'ää'
'I' combining_macron <> 'Ii'
'E' combining_macron <> 'Ee'
'A' combining_macron <> 'Aa'
'O' combining_macron <> 'Oo'
'U' combining_macron <> 'Uu'
'Ä' combining_macron <> 'Ää'


pass(Unicode)

UniClass [VowelL] = ('i' 'e' 'a' 'o' 'u' 'ä')

; all other symbols

cyrillic_small_letter_pe  <> latin_small_letter_p ; 'п'   > 'p'
cyrillic_small_letter_ve  <> latin_small_letter_w ; 'в'   > 'w'
cyrillic_small_letter_em  <> latin_small_letter_m ; 'м'   > 'm'
cyrillic_small_letter_te  <> latin_small_letter_t ; 'т'   > 't'
cyrillic_small_letter_de  <> latin_small_letter_d ; 'д'   > 'd'
cyrillic_small_letter_es  <> latin_small_letter_s ; 'с'   > 's'
cyrillic_small_letter_en  <> latin_small_letter_n ; 'н'   > 'n'
cyrillic_small_letter_el  <> latin_small_letter_l ; 'л'   > 'l'
cyrillic_small_letter_el_with_tail  <> latin_small_letter_l_with_belt ; 'л'   > 'ɬ'  (in Burkova-
TB)
cyrillic_small_letter_che cyrillic_small_letter_soft_sign <> latin_small_letter_c_with_caron ;
'чь'   > 'č'
cyrillic_small_letter_che  <> latin_small_letter_c_with_caron ; 'ч'   > 'č'
cyrillic_small_letter_sha  <> latin_small_letter_s_with_caron ; 'ш'   > 'š'
cyrillic_small_letter_shcha  <> latin_small_letter_s_with_acute ; 'щ'   > 'ś' in Koshkareva 2005
cyrillic_small_letter_short_i  > latin_small_letter_j ; 'й'   > 'j'
cyrillic_small_letter_ka  <> latin_small_letter_k ; 'к'   > 'k'
cyrillic_small_letter_en_with_tail  > latin_small_letter_eng ; 'ң'   > 'ŋ'  (in Burkova-TB)
cyrillic_small_letter_en_with_descender  > latin_small_letter_eng ; 'ң'   > 'ŋ'  (in Burkova-TB)
cyrillic_small_letter_ha  <> latin_small_letter_x ; 'х'   > 'x'
cyrillic_small_letter_soft_sign <> modifier_letter_apostrophe / _ ^[VowelL] ; 'ь'   > ''
cyrillic_small_letter_hard_sign > ; 'ъ' in Koshkareva 2005

cyrillic_capital_letter_pe  <> latin_capital_letter_p ; 'П'   > 'p'
cyrillic_capital_letter_ve  <> latin_capital_letter_w ; 'В'   > 'w'
cyrillic_capital_letter_em  <> latin_capital_letter_m ; 'М'   > 'm'
cyrillic_capital_letter_te  <> latin_capital_letter_t ; 'Т'   > 't'
cyrillic_capital_letter_de  <> latin_capital_letter_d ; 'Д'   > 'd'
cyrillic_capital_letter_es  <> latin_capital_letter_s ; 'С'   > 's'
cyrillic_capital_letter_en  <> latin_capital_letter_n ; 'Н'   > 'n'
cyrillic_capital_letter_el  <> latin_capital_letter_l ; 'Л'   > 'l'
```

```
cyrillic_capital_letter_el_with_tail  <> latin_capital_letter_l_with_belt ; 'д'   > 'Ɬ'  (in
Burkova-TB)
cyrillic_capital_letter_che  cyrillic_capital_letter_soft_sign <>
latin_capital_letter_c_with_caron ; 'чь'   > 'č'
cyrillic_capital_letter_che  <> latin_capital_letter_c_with_caron ; 'ч'   > 'č'
cyrillic_capital_letter_sha  <> latin_capital_letter_s_with_caron ; 'Ш'   > ''
cyrillic_capital_letter_shcha  <> latin_capital_letter_s_with_acute ; 'Щ'   > 'Ś' in Koshkareva
2005
cyrillic_capital_letter_short_i  > latin_capital_letter_j ; 'й'   > 'j'
cyrillic_capital_letter_ka  <> latin_capital_letter_k ; 'к'   > 'k'
cyrillic_capital_letter_en_with_tail  > latin_capital_letter_eng ; 'ң'   > 'ŋ'  (in Burkova-TB)
cyrillic_capital_letter_en_with_descender  > latin_capital_letter_eng ; 'Ң'   > 'ŋ'  (in Burkova-
TB)
cyrillic_capital_letter_ha  <> latin_capital_letter_x ; 'х'   > 'x'
cyrillic_capital_letter_soft_sign > modifier_letter_apostrophe; 'ь'   > '’'
cyrillic_capital_letter_hard_sign > ; 'ъ' in Koshkareva 2005

; only in Russian loans

cyrillic_small_letter_be  <> latin_small_letter_b ; 'б'   > 'b'
cyrillic_capital_letter_be  <> latin_capital_letter_b ; 'б'   > 'b'
cyrillic_small_letter_ghe <> latin_small_letter_g ; 'г'   > 'g'
cyrillic_capital_letter_ghe  <> latin_capital_letter_g ; 'г'   > 'g'
cyrillic_small_letter_ef  <> latin_small_letter_f ; 'ф'   > ''
cyrillic_capital_letter_ef  <> latin_capital_letter_f ; 'Ф'   > ''
cyrillic_small_letter_ze  <> latin_small_letter_z ; 'з'   > ''
cyrillic_capital_letter_ze  <> latin_capital_letter_z ; 'З'   > ''
cyrillic_small_letter_er  <> latin_small_letter_r ; 'р'   > 'r'
cyrillic_capital_letter_er  <> latin_capital_letter_r ; 'р'   > 'r'
cyrillic_small_letter_zhe  <> latin_small_letter_z_with_caron ; 'ж'   > ''
cyrillic_capital_letter_zhe  <> latin_capital_letter_z_with_caron ; 'Ж'   > ''
cyrillic_small_letter_tse  <> latin_small_letter_c ; 'ц'   > ''
cyrillic_capital_letter_tse  <> latin_capital_letter_c ; 'Ц'   > ''


pass(Unicode)

'«' > '“'
'»' > '”'
```

Table 21. Transliteration rules for Burkova 2012

```
pass(Unicode)

UniClass [VowelNonPalC] =     ('ы' 'э' 'а' 'о' 'у') ; 'ы''э''а''о''у'
UniClass [VowelL] =          ('i' 'e' 'a' 'o' 'u') ; 'i''e''a''o''u'
UniClass [VowelNonPalCCap] = ('Ы' 'Э' 'А' 'О' 'У') ; 'Ы''Э''А''О''У'
UniClass [VowelLCap] =         ('I' 'E' 'A' 'O' 'U') ; 'I''E''A''O''U'

; glottal stop

'”' <> latin_letter_glottal_stop ; U+201D (in Burkova-TB)
'’' <> latin_letter_glottal_stop ; U+2019 (in Burkova-TB)

; cyrillic vowels with macron

cyrillic_small_letter_i_with_macron <> cyrillic_small_letter_i combining_macron ;й
cyrillic_small_letter_u_with_macron <> cyrillic_small_letter_u combining_macron ;ў
cyrillic_capital_letter_i_with_macron <> cyrillic_capital_letter_i combining_macron ;Й
cyrillic_capital_letter_u_with_macron <> cyrillic_capital_letter_u combining_macron ;Ў
cyrillic_small_letter_short_u <> cyrillic_small_letter_u combining_breve ;ў
cyrillic_capital_letter_short_u <> cyrillic_capital_letter_u combining_breve ;Ў


; plain vowels reverse direction

[VowelNonPalC] < [VowelL]
[VowelNonPalCCap] < [VowelLCap]

pass(Unicode)

; и intervocalically

UniClass [VowelC] = ('ы' 'э' 'а' 'о' 'у' 'и' 'е' 'я' 'ё' 'ю' 'ä' 'Ы' 'Э' 'А' 'О' 'У' 'И' 'Е' 'Я'
'Ё' 'Ю' 'Ä')

[VowelC] cyrillic_small_letter_i <> [VowelC] latin_small_letter_j latin_small_letter_i


pass(Unicode)

; vowels after D, N (palat.) and other vowels

UniClass [ConsPalC] = ('п' 'б' 'в' 'м' 'т' 'д' 'н' 'л' 'р' 'ц' 'с' 'з' 'к' 'х' 'ң' 'ң') ; 'ң' (in
Burkova-TB)
UniClass [ConsPalCCap] = ('П' 'Б' 'В' 'М' 'Т' 'Д' 'Н' 'Л' 'Р' 'Ц' 'С' 'З' 'К' 'Х' 'Ң' 'Ң') ; 'Ң'
(in Burkova-TB)
UniClass [ConsPalL] = ('p' 'b' 'w' 'm' 't' 'd' 'n' 'l' 'r' 'c' 's' 'z' 'k' 'x' 'ŋ' 'ŋ') ;
UniClass [ConsPalLCap] = ('P' 'B' 'W' 'M' 'T' 'D' 'N' 'L' 'R' 'C' 'S' 'Z' 'K' 'X' 'Ŋ' 'Ŋ') ;

UniClass [ConsNonPalC] = ('ч''ш') ; ч ш
UniClass [ConsNonPalCCap] = ('Ч''Ш') ; Ч Ш
UniClass [ConsNonPalL] = ('č''š') ; č š
UniClass [ConsNonPalLCap] = ('Č''Š') ; Č Š


UniClass [VowelPalC] =        ('и' 'е' 'я' 'ё' 'ю') ;
UniClass [VowelNonPalC] =     ('ы' 'э' 'а' 'о' 'у') ;
UniClass [VowelL] =          ('i' 'e' 'a' 'o' 'u') ;
UniClass [VowelPalXiC] =      ('e' 'я' 'ё' 'ю') ;
UniClass [VowelXiL] =  ('e' 'a' 'o' 'u') ;

UniClass [VowelPalCCap] =       ('И' 'Е' 'Я' 'Ё' 'Ю') ;
UniClass [VowelNonPalCCap] = ('Ы' 'Э' 'А' 'О' 'У') ;
UniClass [VowelLCap] =         ('I' 'E' 'A' 'O' 'U') ;
UniClass [VowelPalXiCCap] =  ('E' 'Я' 'Ё' 'Ю') ;
UniClass [VowelXiLCap] =       ('E' 'A' 'O' 'U') ;


[ConsPalC]=c [VowelPalC]=v > [ConsPalL]=c modifier_letter_apostrophe [VowelL]=v
[ConsNonPalC]=c [VowelPalC]=v > [ConsNonPalL]=c [VowelL]=v
[VowelPalC]=v <> latin_small_letter_j [VowelL]=v
[VowelNonPalC] > [VowelL]
```

```
[ConsPalCCap]=c [VowelPalC]=v > [ConsPalLCap]=c modifier_letter_apostrophe [VowelL]=v
[ConsNonPalCCap]=c [VowelPalC]=v > [ConsNonPalLCap]=c [VowelL]=v
[VowelPalCCap]=v <> latin_capital_letter_j [VowelL]=v
[VowelNonPalCCap] > [VowelLCap]

[VowelPalC]=v < modifier_letter_apostrophe [VowelL]=v
[VowelPalCCap]=v < modifier_letter_apostrophe [VowelLCap]=v


pass(Unicode)

; long vowels

'i' combining_macron <> 'ii'
'e' combining_macron <> 'ee'
'a' combining_macron <> 'aa'
'o' combining_macron <> 'oo'
'u' combining_macron <> 'uu'
'ä' combining_macron <> 'ää'
'I' combining_macron <> 'Ii'
'E' combining_macron <> 'Ee'
'A' combining_macron <> 'Aa'
'O' combining_macron <> 'Oo'
'U' combining_macron <> 'Uu'
'Ä' combining_macron <> 'Ää'


pass(Unicode)

UniClass [VowelL] = ('i' 'e' 'a' 'o' 'u' 'ä')

; all other symbols

cyrillic_small_letter_pe  <> latin_small_letter_p ; 'п'   > 'p'
cyrillic_small_letter_ve  <> latin_small_letter_w ; 'в'   > 'w'
cyrillic_small_letter_em  <> latin_small_letter_m ; 'м'   > 'm'
cyrillic_small_letter_te  <> latin_small_letter_t ; 'т'   > 't'
cyrillic_small_letter_de  <> latin_small_letter_d ; 'д'   > 'd'
cyrillic_small_letter_es  <> latin_small_letter_s ; 'с'   > 's'
cyrillic_small_letter_en  <> latin_small_letter_n ; 'н'   > 'n'
cyrillic_small_letter_el  <> latin_small_letter_l ; 'л'   > 'l'
cyrillic_small_letter_tse  <> latin_small_letter_c ; 'ц'   > ''
cyrillic_small_letter_che cyrillic_small_letter_soft_sign <> latin_small_letter_c_with_caron ;
'чь'   > 'č'
cyrillic_small_letter_che  <> latin_small_letter_c_with_caron ; 'ч'   > 'č'
cyrillic_small_letter_sha  <> latin_small_letter_s_with_caron ; 'ш'   > 'š'
cyrillic_small_letter_shcha  <> latin_small_letter_s_with_acute ; 'щ'   > 'ś' in Koshkareva 2005
cyrillic_small_letter_short_i  > latin_small_letter_j ; 'й'   > 'j'
cyrillic_small_letter_ka  <> latin_small_letter_k ; 'к'   > 'k'
cyrillic_small_letter_en_with_tail  > latin_small_letter_eng ; 'ң'   > 'ŋ'  (in Burkova-TB)
cyrillic_small_letter_en_with_descender  > latin_small_letter_eng ; 'ң'   > 'ŋ'  (in Burkova-TB)
cyrillic_small_letter_ha  <> latin_small_letter_x ; 'х'   > 'x'
cyrillic_small_letter_soft_sign <> modifier_letter_apostrophe / _ ^[VowelL] ; 'ь'   > ''''
cyrillic_small_letter_hard_sign > ; 'ъ' in Koshkareva 2005

cyrillic_capital_letter_pe  <> latin_capital_letter_p ; 'п'   > 'p'
cyrillic_capital_letter_ve  <> latin_capital_letter_w ; 'в'   > 'w'
cyrillic_capital_letter_em  <> latin_capital_letter_m ; 'м'   > 'm'
cyrillic_capital_letter_te  <> latin_capital_letter_t ; 'т'   > 't'
cyrillic_capital_letter_de  <> latin_capital_letter_d ; 'д'   > 'd'
cyrillic_capital_letter_es  <> latin_capital_letter_s ; 'с'   > 's'
cyrillic_capital_letter_en  <> latin_capital_letter_n ; 'н'   > 'n'
cyrillic_capital_letter_el  <> latin_capital_letter_l ; 'л'   > 'l'
cyrillic_capital_letter_tse  <> latin_capital_letter_c ; 'Ц'   > ''
cyrillic_capital_letter_che  cyrillic_capital_letter_soft_sign <>
latin_capital_letter_c_with_caron ; 'чь'   > 'č'
cyrillic_capital_letter_che  <> latin_capital_letter_c_with_caron ; 'ч'   > 'č'
cyrillic_capital_letter_sha  <> latin_capital_letter_s_with_caron ; 'Ш'   > ''
cyrillic_capital_letter_shcha  <> latin_capital_letter_s_with_acute ; 'Щ'   > 'Ś' in Koshkareva
2005
cyrillic_capital_letter_short_i  > latin_capital_letter_j ; 'й'   > 'j'
cyrillic_capital_letter_ka  <> latin_capital_letter_k ; 'к'   > 'k'
```

```
cyrillic_capital_letter_en_with_tail  > latin_capital_letter_eng ; 'ҥ'   > 'ŋ'  (in Burkova-TB)
cyrillic_capital_letter_en_with_descender  > latin_capital_letter_eng ; 'Ҥ'   > 'ŋ'  (in Burkova-
TB)
cyrillic_capital_letter_ha  <> latin_capital_letter_x ; 'х'   > 'x'
cyrillic_capital_letter_soft_sign > modifier_letter_apostrophe; 'Ь'   > 'ʼ'
cyrillic_capital_letter_hard_sign > ; 'ъ' in Koshkareva 2005

; only in Russian loans

cyrillic_small_letter_be  <> latin_small_letter_b ; 'б'   > 'b'
cyrillic_capital_letter_be  <> latin_capital_letter_b ; 'б'   > 'b'
cyrillic_small_letter_ghe <> latin_small_letter_g ; 'г'   > 'g'
cyrillic_capital_letter_ghe  <> latin_capital_letter_g ; 'г'   > 'g'
cyrillic_small_letter_ef  <> latin_small_letter_f ; 'ф'   > ''
cyrillic_capital_letter_ef  <> latin_capital_letter_f ; 'Ф'   > ''
cyrillic_small_letter_ze  <> latin_small_letter_z ; 'з'   > ''
cyrillic_capital_letter_ze  <> latin_capital_letter_z ; 'З'   > ''
cyrillic_small_letter_er  <> latin_small_letter_r ; 'р'   > 'r'
cyrillic_capital_letter_er  <> latin_capital_letter_r ; 'р'   > 'r'
cyrillic_small_letter_zhe  <> latin_small_letter_z_with_caron ; 'ж'   > ''
cyrillic_capital_letter_zhe  <> latin_capital_letter_z_with_caron ; 'Ж'   > ''


pass(Unicode)

'«' > '‟'
'»' > '‟'
```

Table 22. Transliteration rules for Koshkareva 2005

```
pass(Unicode)

UniClass [VowelNonPalC] =     ('ы' 'э' 'а' 'о' 'у') ; 'ы''э''а''о''у'
UniClass [VowelL] =          ('i' 'e' 'a' 'o' 'u') ; 'i''e''a''o''u'
UniClass [VowelNonPalCCap] = ('Ы' 'Э' 'А' 'О' 'У') ; 'Ы''Э''А''О''У'
UniClass [VowelLCap] =         ('I' 'E' 'A' 'O' 'U') ; 'I''E''A''O''U'

; glottal stop

'"' <> latin_letter_glottal_stop ;  (in Koshkareva-2005)
apostrophe > latin_letter_glottal_stop ; ' (in Koshkareva-2005)
'’’' > latin_letter_glottal_stop
'”' > latin_letter_glottal_stop
'’' > latin_letter_glottal_stop

; diphthong

'э' <> 'ä' ; э'□' (U+F50F) in the Word file, looks like 'э '
'э□' > 'ä' ; '□' U+F50F, looks like 'э_'
'э_' > 'ä' ; '_', looks like 'э_'
'Э' <> 'Ä' ; э'□' (U+F50F) in the Word file, looks like 'э '
'Э□' > 'Ä' ; '□' U+F50F, looks like 'э_'
'Э_' > 'Ä' ; '_', looks like 'э_'

; plain vowels reverse direction

[VowelNonPalC] < [VowelL]
[VowelNonPalCCap] < [VowelLCap]

pass(Unicode)

; и intervocalically

UniClass [VowelC] = ('ы' 'э' 'а' 'о' 'у' 'и' 'е' 'я' 'ё' 'ю' 'ä' 'Ы' 'Э' 'А' 'О' 'У' 'И' 'Е' 'Я'
'Ё' 'Ю' 'Ä')

[VowelC] cyrillic_small_letter_i <> [VowelC] latin_small_letter_j latin_small_letter_i


pass(Unicode)

; vowels after D, N (palat.) and other vowels

UniClass [ConsPalC] = ('п' 'в' 'м' 'т' 'д' 'н' 'л' 'ӆ' 'к' 'х' 'ӈ') ;
UniClass [ConsPalCCap] = ('П' 'В' 'М' 'Т' 'Д' 'Н' 'Л' 'Ӆ' 'К' 'Х' 'Ӈ') ;
UniClass [ConsPalL] = ('p' 'w' 'm' 't' 'd' 'n' 'l' 'ɬ' 'k' 'x' 'ŋ') ;
UniClass [ConsPalLCap] = ('P' 'W' 'M' 'T' 'D' 'N' 'L' 'Ɬ' 'K' 'X' 'Ŋ') ;

UniClass [ConsNonPalC] = ('ч''ш''с') ; ч ш
UniClass [ConsNonPalCCap] = ('Ч''Ш''С') ; Ч Ш
UniClass [ConsNonPalL] = ('č''š''s') ; č š
UniClass [ConsNonPalLCap] = ('Č''Š''S') ; Č Š


UniClass [VowelPalC] =        ('и' 'е' 'я' 'ё' 'ю') ;
UniClass [VowelNonPalC] =     ('ы' 'э' 'а' 'о' 'у') ;
UniClass [VowelL] =          ('i' 'e' 'a' 'o' 'u') ;
UniClass [VowelPalXiC] =      ('e' 'я' 'ё' 'ю') ;
UniClass [VowelXiL] =  ('e' 'a' 'o' 'u') ;

UniClass [VowelPalCCap] =       ('И' 'Е' 'Я' 'Ё' 'Ю') ;
UniClass [VowelNonPalCCap] = ('Ы' 'Э' 'А' 'О' 'У') ;
UniClass [VowelLCap] =         ('I' 'E' 'A' 'O' 'U') ;
UniClass [VowelPalXiCCap] =  ('E' 'Я' 'Ё' 'Ю') ;
UniClass [VowelXiLCap] =       ('E' 'A' 'O' 'U') ;


[ConsPalC]=c [VowelPalC]=v > [ConsPalL]=c modifier_letter_apostrophe [VowelL]=v
[ConsNonPalC]=c [VowelPalC]=v > [ConsNonPalL]=c [VowelL]=v
[VowelPalC]=v <> latin_small_letter_j [VowelL]=v
[VowelNonPalC] > [VowelL]
```

```
[ConsPalCCap]=c [VowelPalC]=v > [ConsPalLCap]=c modifier_letter_apostrophe [VowelL]=v
[ConsNonPalCCap]=c [VowelPalC]=v > [ConsNonPalLCap]=c [VowelL]=v
[VowelPalCCap]=v <> latin_capital_letter_j [VowelL]=v
[VowelNonPalCCap] > [VowelLCap]

[VowelPalC]=v < modifier_letter_apostrophe [VowelL]=v
[VowelPalCCap]=v < modifier_letter_apostrophe [VowelLCap]=v


pass(Unicode)

; long vowels

'i' combining_macron <> 'ii'
'e' combining_macron <> 'ee'
'a' combining_macron <> 'aa'
'o' combining_macron <> 'oo'
'u' combining_macron <> 'uu'
'ä' combining_macron <> 'ää'
'I' combining_macron <> 'Ii'
'E' combining_macron <> 'Ee'
'A' combining_macron <> 'Aa'
'O' combining_macron <> 'Oo'
'U' combining_macron <> 'Uu'
'Ä' combining_macron <> 'Ää'


pass(Unicode)

UniClass [VowelL] = ('i' 'e' 'a' 'o' 'u' 'ä')

; all other symbols

cyrillic_small_letter_pe  <> latin_small_letter_p ; 'п'   > 'p'
cyrillic_small_letter_ve  <> latin_small_letter_w ; 'в'   > 'w'
cyrillic_small_letter_em  <> latin_small_letter_m ; 'м'   > 'm'
cyrillic_small_letter_te  <> latin_small_letter_t ; 'т'   > 't'
cyrillic_small_letter_de  <> latin_small_letter_d ; 'д'   > 'd'
cyrillic_small_letter_es  <> latin_small_letter_s ; 'с'   > 's'
cyrillic_small_letter_en  <> latin_small_letter_n ; 'н'   > 'n'
cyrillic_small_letter_el  <> latin_small_letter_l ; 'л'   > 'l'
cyrillic_small_letter_el_with_hook  <> latin_small_letter_l_with_belt ; 'ԯ'   > 'ɬ'
cyrillic_small_letter_che cyrillic_small_letter_soft_sign <> latin_small_letter_c_with_caron ;
'чь'   > 'č'
cyrillic_small_letter_che  <> latin_small_letter_c_with_caron ; 'ч'   > 'č'
cyrillic_small_letter_sha  <> latin_small_letter_s_with_caron ; 'ш'   > 'š'
cyrillic_small_letter_shcha  <> latin_small_letter_s_with_acute ; 'щ'   > 'ś' in Koshkareva 2005
cyrillic_small_letter_short_i  > latin_small_letter_j ; 'й'   > 'j'
cyrillic_small_letter_ka  <> latin_small_letter_k ; 'к'   > 'k'
cyrillic_small_letter_en_with_hook  > latin_small_letter_eng ; 'ԋ'   > 'ŋ'
cyrillic_small_letter_en_with_descender  <> latin_small_letter_eng ; 'ҥ'   > 'ŋ'
cyrillic_small_letter_ha  <> latin_small_letter_x ; 'х'   > 'x'
cyrillic_small_letter_soft_sign <> modifier_letter_apostrophe / _ ^[VowelL] ; 'ь'   > ''
cyrillic_small_letter_hard_sign > ; 'ъ' in Koshkareva 2005

cyrillic_capital_letter_pe  <> latin_capital_letter_p ; 'П'   > 'p'
cyrillic_capital_letter_ve  <> latin_capital_letter_w ; 'В'   > 'w'
cyrillic_capital_letter_em  <> latin_capital_letter_m ; 'М'   > 'm'
cyrillic_capital_letter_te  <> latin_capital_letter_t ; 'Т'   > 't'
cyrillic_capital_letter_de  <> latin_capital_letter_d ; 'Д'   > 'd'
cyrillic_capital_letter_es  <> latin_capital_letter_s ; 'С'   > 's'
cyrillic_capital_letter_en  <> latin_capital_letter_n ; 'Н'   > 'n'
cyrillic_capital_letter_el  <> latin_capital_letter_l ; 'Л'   > 'l'
cyrillic_capital_letter_el_with_hook  <> latin_capital_letter_l_with_belt ; 'Ԯ'   > 'ɬ'
cyrillic_capital_letter_che  cyrillic_capital_letter_soft_sign <>
latin_capital_letter_c_with_caron ; 'чь'   > 'č'
cyrillic_capital_letter_che  <> latin_capital_letter_c_with_caron ; 'Ч'   > 'č'
cyrillic_capital_letter_sha  <> latin_capital_letter_s_with_caron ; 'Ш'   > ''
cyrillic_capital_letter_shcha  <> latin_capital_letter_s_with_acute ; 'Щ'   > 'Ś' in Koshkareva
2005
cyrillic_capital_letter_short_i  > latin_capital_letter_j ; 'Й'   > 'j'
cyrillic_capital_letter_ka  <> latin_capital_letter_k ; 'К'   > 'k'
```

```
cyrillic_capital_letter_en_with_hook  > latin_capital_letter_eng ; 'Ҥ'   > 'ŋ'
cyrillic_capital_letter_en_with_descender  <> latin_capital_letter_eng ; 'Ҥ'   > 'ŋ'
cyrillic_capital_letter_ha  <> latin_capital_letter_x ; 'х'   > 'x'
cyrillic_capital_letter_soft_sign > modifier_letter_apostrophe; 'Ь'   > 'ʼ'
cyrillic_capital_letter_hard_sign > ; 'ъ' in Koshkareva 2005

; only in Russian loans

cyrillic_small_letter_be  <> latin_small_letter_b ; 'б'   > 'b'
cyrillic_capital_letter_be  <> latin_capital_letter_b ; 'б'   > 'b'
cyrillic_small_letter_ghe <> latin_small_letter_g ; 'г'   > 'g'
cyrillic_capital_letter_ghe  <> latin_capital_letter_g ; 'г'   > 'g'
cyrillic_small_letter_ef  <> latin_small_letter_f ; 'ф'   > ''
cyrillic_capital_letter_ef  <> latin_capital_letter_f ; 'Ф'   > ''
cyrillic_small_letter_ze  <> latin_small_letter_z ; 'з'   > ''
cyrillic_capital_letter_ze  <> latin_capital_letter_z ; 'З'   > ''
cyrillic_small_letter_er  <> latin_small_letter_r ; 'р'   > 'r'
cyrillic_capital_letter_er  <> latin_capital_letter_r ; 'р'   > 'r'
cyrillic_small_letter_zhe  <> latin_small_letter_z_with_caron ; 'ж'   > ''
cyrillic_capital_letter_zhe  <> latin_capital_letter_z_with_caron ; 'Ж'   > ''
cyrillic_small_letter_tse  <> latin_small_letter_c ; 'ц'   > ''
cyrillic_capital_letter_tse  <> latin_capital_letter_c ; 'Ц'   > ''


pass(Unicode)

'«' > '"'
'»' > '"'
```

Table 23. Transliteration for Koshkareva et al. 2007

```
pass(Unicode)

UniClass [VowelNonPalC] =     ('ы' 'э' 'а' 'о' 'у') ; 'ы''э''а''о''у'
UniClass [VowelL] =          ('i' 'e' 'a' 'o' 'u') ; 'i''e''a''o''u'
UniClass [VowelNonPalCCap] = ('Ы' 'Э' 'А' 'О' 'У') ; 'Ы''Э''А''О''У'
UniClass [VowelLCap] =         ('I' 'E' 'A' 'O' 'U') ; 'I''E''A''O''U'

; glottal stop
"'"> latin_letter_glottal_stop
'"' <> latin_letter_glottal_stop
apostrophe <> latin_letter_glottal_stop ;'U+0027' <> latin_letter_glottal_stop

; diphthong

'э(' > 'ä'
'э˘' > 'ĕ'
; plain vowels reverse direction

[VowelNonPalC] < [VowelL]
[VowelNonPalCCap] < [VowelLCap]

pass(Unicode)

; и intervocalically

UniClass [VowelC] = ('ы' 'э' 'а' 'о' 'у' 'и' 'е' 'я' 'ё' 'ю' 'ä')

[VowelC] cyrillic_small_letter_i <> [VowelC] latin_small_letter_j latin_small_letter_i


pass(Unicode)

; vowels after D, N (palat.) and other vowels

UniClass [ConsPalC] = ('п' 'в' 'м' 'т' 'д' 'н' 'л' 'ӆ' 'к' 'х' 'ӈ') ;
UniClass [ConsPalCCap] = ('П' 'В' 'М' 'Т' 'Д' 'Н' 'Л' 'Ӆ' 'К' 'Х' 'Ӈ') ;
UniClass [ConsPalL] = ('p' 'w' 'm' 't' 'd' 'n' 'l' 'ɬ' 'k' 'x' 'ŋ') ;
UniClass [ConsPalLCap] = ('P' 'W' 'M' 'T' 'D' 'N' 'L' 'Ɬ' 'K' 'X' 'Ŋ') ;

UniClass [ConsNonPalC] = ('ч''ш''с') ; ч ш
UniClass [ConsNonPalCCap] = ('Ч''Ш''С') ; Ч Ш
UniClass [ConsNonPalL] = ('č''š''s') ; č š
UniClass [ConsNonPalLCap] = ('Č''Š''S') ; Č Š


UniClass [VowelPalC] =        ('и' 'е' 'я' 'ё' 'ю') ;
UniClass [VowelNonPalC] =     ('ы' 'э' 'а' 'о' 'у') ;
UniClass [VowelL] =          ('i' 'e' 'a' 'o' 'u') ;
UniClass [VowelPalXiC] =      ('е' 'я' 'ё' 'ю') ;
UniClass [VowelXiL] =  ('e' 'a' 'o' 'u') ;

UniClass [VowelPalCCap] =       ('И' 'Е' 'Я' 'Ё' 'Ю') ;
UniClass [VowelNonPalCCap] = ('Ы' 'Э' 'А' 'О' 'У') ;
UniClass [VowelLCap] =          ('I' 'E' 'A' 'O' 'U') ;
UniClass [VowelPalXiCCap] =  ('Е' 'Я' 'Ё' 'Ю') ;
UniClass [VowelXiLCap] =        ('E' 'A' 'O' 'U') ;


[ConsPalC]=c [VowelPalC]=v > [ConsPalL]=c modifier_letter_apostrophe [VowelL]=v
[ConsNonPalC]=c [VowelPalC]=v > [ConsNonPalL]=c [VowelL]=v
[VowelPalC]=v <> latin_small_letter_j [VowelL]=v
[VowelNonPalC] > [VowelL]

[ConsPalCCap]=c [VowelPalC]=v > [ConsPalLCap]=c modifier_letter_apostrophe [VowelL]=v
[ConsNonPalCCap]=c [VowelPalC]=v > [ConsNonPalLCap]=c [VowelL]=v
[VowelPalCCap]=v <> latin_capital_letter_j [VowelL]=v
[VowelNonPalCCap] > [VowelLCap]

[VowelPalC]=v < modifier_letter_apostrophe [VowelL]=v
[VowelPalCCap]=v < modifier_letter_apostrophe [VowelLCap]=v
```

```
pass(Unicode)

; long vowels

'i' combining_macron <> 'ii'
'e' combining_macron <> 'ee'
'a' combining_macron <> 'aa'
'o' combining_macron <> 'oo'
'u' combining_macron <> 'uu'
'ä' combining_macron <> 'ää'
'I' combining_macron <> 'Ii'
'E' combining_macron <> 'Ee'
'A' combining_macron <> 'Aa'
'O' combining_macron <> 'Oo'
'U' combining_macron <> 'Uu'
'Ä' combining_macron <> 'Ää'


pass(Unicode)

UniClass [VowelL] = ('i' 'e' 'a' 'o' 'u' 'ä')

; all other symbols

cyrillic_small_letter_pe  <> latin_small_letter_p ; 'п'   > 'p'
cyrillic_small_letter_ve  <> latin_small_letter_w ; 'в'   > 'w'
cyrillic_small_letter_em  <> latin_small_letter_m ; 'м'   > 'm'
cyrillic_small_letter_te  <> latin_small_letter_t ; 'т'   > 't'
cyrillic_small_letter_de  <> latin_small_letter_d ; 'д'   > 'd'
cyrillic_small_letter_es  <> latin_small_letter_s ; 'с'   > 's'
cyrillic_small_letter_en  <> latin_small_letter_n ; 'н'   > 'n'
cyrillic_small_letter_el  <> latin_small_letter_l ; 'л'   > 'l'
cyrillic_small_letter_el_with_hook  <> latin_small_letter_l_with_belt ; 'ԯ'   > 'ɬ'
cyrillic_small_letter_che cyrillic_small_letter_soft_sign <> latin_small_letter_c_with_caron ;
'чь'   > 'č'
cyrillic_small_letter_che  <> latin_small_letter_c_with_caron ; 'ч'   > 'č'
cyrillic_small_letter_sha  <> latin_small_letter_s_with_caron ; 'ш'   > ''
cyrillic_small_letter_short_i  > latin_small_letter_j ; 'й'   > 'j'
cyrillic_small_letter_ka  <> latin_small_letter_k ; 'к'   > 'k'
cyrillic_small_letter_en_with_hook  > latin_small_letter_eng ; 'ԋ'   > 'ŋ'
cyrillic_small_letter_en_with_descender  <> latin_small_letter_eng ; 'ԧ'   > 'ŋ'
cyrillic_small_letter_ha  <> latin_small_letter_x ; 'х'   > 'x'
cyrillic_small_letter_soft_sign <> modifier_letter_apostrophe / _ ^[VowelL] ; 'ь'   > '’'

cyrillic_capital_letter_pe  <> latin_capital_letter_p ; 'П'   > 'p'
cyrillic_capital_letter_ve  <> latin_capital_letter_w ; 'В'   > 'w'
cyrillic_capital_letter_em  <> latin_capital_letter_m ; 'М'   > 'm'
cyrillic_capital_letter_te  <> latin_capital_letter_t ; 'Т'   > 't'
cyrillic_capital_letter_de  <> latin_capital_letter_d ; 'Д'   > 'd'
cyrillic_capital_letter_es  <> latin_capital_letter_s ; 'С'   > 's'
cyrillic_capital_letter_en  <> latin_capital_letter_n ; 'Н'   > 'n'
cyrillic_capital_letter_el  <> latin_capital_letter_l ; 'Л'   > 'l'
cyrillic_capital_letter_el_with_hook  <> latin_capital_letter_l_with_belt ; 'ԯ'   > 'ɬ'
'л' > 'ɬ'
'Л' > 'ɬ'
cyrillic_capital_letter_che  cyrillic_capital_letter_soft_sign <>
latin_capital_letter_c_with_caron ; 'чь'   > 'č'
cyrillic_capital_letter_che  <> latin_capital_letter_c_with_caron ; 'Ч'   > 'č'
cyrillic_capital_letter_sha  <> latin_capital_letter_s_with_caron ; 'Ш'   > ''
cyrillic_capital_letter_short_i  > latin_capital_letter_j ; 'Й'   > 'j'
cyrillic_capital_letter_ka  <> latin_capital_letter_k ; 'К'   > 'k'
cyrillic_capital_letter_en_with_hook  > latin_capital_letter_eng ; 'Ԋ'   > 'ŋ'
cyrillic_capital_letter_en_with_descender  <> latin_capital_letter_eng ; 'Ԧ'   > 'ŋ'
cyrillic_capital_letter_ha  <> latin_capital_letter_x ; 'Х'   > 'x'
cyrillic_capital_letter_soft_sign > modifier_letter_apostrophe; 'Ь'   > '’'

; only in Russian loans

cyrillic_small_letter_be  <> latin_small_letter_b ; 'б'   > 'b'
cyrillic_capital_letter_be  <> latin_capital_letter_b ; 'Б'   > 'b'
cyrillic_small_letter_ghe <> latin_small_letter_g ; 'г'   > 'g'
cyrillic_capital_letter_ghe  <> latin_capital_letter_g ; 'Г'   > 'g'
```

```
cyrillic_small_letter_ef  <> latin_small_letter_f ; 'ф'   > ''
cyrillic_capital_letter_ef  <> latin_capital_letter_f ; 'Ф'   > ''
cyrillic_small_letter_ze  <> latin_small_letter_z ; 'з'   > ''
cyrillic_capital_letter_ze  <> latin_capital_letter_z ; 'З'   > ''
cyrillic_small_letter_er  <> latin_small_letter_r ; 'р'   > 'r'
cyrillic_capital_letter_er  <> latin_capital_letter_r ; 'р'   > 'r'
cyrillic_small_letter_zhe  <> latin_small_letter_z_with_caron ; 'ж'   > ''
cyrillic_capital_letter_zhe  <> latin_capital_letter_z_with_caron ; 'Ж'   > ''
cyrillic_small_letter_tse  <> latin_small_letter_c ; 'ц'   > ''
cyrillic_capital_letter_tse  <> latin_capital_letter_c ; 'Ц'   > ''
cyrillic_small_letter_shcha  <> latin_small_letter_s_with_caron latin_small_letter_c_with_caron ;
'щ'   > ''
cyrillic_capital_letter_shcha  <> latin_capital_letter_s_with_caron
latin_small_letter_c_with_caron ; 'Щ'   > ''

pass(Unicode)

'«' > '"'
'»' > '"'
```

Table 24. Transliteration rules for Labanauskas 2001

```
pass(Unicode)

UniClass [VowelNonPalC] =     ('ы' 'э' 'а' 'о' 'у') ; 'ы''э''а''о''у'
UniClass [VowelL] =          ('i' 'e' 'a' 'o' 'u') ; 'i''e''a''o''u'
UniClass [VowelNonPalCCap] = ('Ы' 'Э' 'А' 'О' 'У') ; 'Ы''Э''А''О''У'
UniClass [VowelLCap] =         ('I' 'E' 'A' 'O' 'U') ; 'I''E''A''O''U'

; glottal stop

'"' <> latin_letter_glottal_stop
U+0027 > latin_letter_glottal_stop

; plain vowels reverse direction

[VowelNonPalC] < [VowelL]
[VowelNonPalCCap] < [VowelLCap]

pass(Unicode)

; и intervocalically

UniClass [VowelC] = ('ы' 'э' 'а' 'о' 'у' 'и' 'е' 'я' 'ё' 'ю')

[VowelC] cyrillic_small_letter_i <> [VowelC] latin_small_letter_j latin_small_letter_i


pass(Unicode)

; vowels after D, N (palat.) and other vowels

UniClass [ConsPalC] = ('т' 'д' 'н' 'с' 'л' 'р' 'п' 'б' 'м' 'х') ; т д н с л р п б м х
UniClass [ConsPalCCap] = ('Т' 'Д' 'Н' 'С' 'Л' 'Р' 'П' 'Б' 'М' 'Х') ; Т Д Н С Л Р П Б М Х
UniClass [ConsPalL] = ('t' 'd' 'n' 's' 'l' 'r' 'p' 'b' 'm' 'x') ; t d n s l r p b m x
UniClass [ConsPalLCap] = ('T' 'D' 'N' 'S' 'L' 'R' 'P' 'B' 'M' 'X') ; T D N S L R P B M X

UniClass [ConsNonPalC] = ('ч') ; ч
UniClass [ConsNonPalCCap] = ('Ч') ; Ч
UniClass [ConsNonPalL] = ('č') ; č
UniClass [ConsNonPalLCap] = ('Č') ; Č


UniClass [VowelPalC] =        ('и' 'е' 'я' 'ё' 'ю') ; 'и''е''я''ё''ю'
UniClass [VowelNonPalC] =     ('ы' 'э' 'а' 'о' 'у') ; 'ы''э''а''о''у'
UniClass [VowelL] =          ('i' 'e' 'a' 'o' 'u') ; 'i''e''a''o''u'
UniClass [VowelPalXiC] =      ('е' 'я' 'ё' 'ю') ; 'е''я''ё''ю'
UniClass [VowelXiL] =  ('e' 'a' 'o' 'u') ; 'e''a''o''u'

UniClass [VowelPalCCap] =       ('И' 'Е' 'Я' 'Ё' 'Ю') ; 'И''Е''Я''Ё''Ю'
UniClass [VowelNonPalCCap] = ('Ы' 'Э' 'А' 'О' 'У') ; 'Ы''Э''А''О''У'
UniClass [VowelLCap] =         ('I' 'E' 'A' 'O' 'U') ; 'I''E''A''O''U'
UniClass [VowelPalXiCCap] =  ('Е' 'Я' 'Ё' 'Ю') ; 'Е''Я''Ё''Ю'
UniClass [VowelXiLCap] =       ('E' 'A' 'O' 'U') ; 'E''A''O''U'


[ConsPalC]=c [VowelPalC]=v > [ConsPalL]=c modifier_letter_apostrophe [VowelL]=v
[ConsNonPalC]=c [VowelPalC]=v > [ConsNonPalL]=c [VowelL]=v
[VowelPalC]=v <> latin_small_letter_j [VowelL]=v
[VowelNonPalC] > [VowelL]

[ConsPalCCap]=c [VowelPalC]=v > [ConsPalLCap]=c modifier_letter_apostrophe [VowelL]=v
[ConsNonPalCCap]=c [VowelPalC]=v > [ConsNonPalLCap]=c [VowelL]=v
[VowelPalCCap]=v <> latin_capital_letter_j [VowelL]=v
[VowelNonPalCCap] > [VowelLCap]

[VowelPalC]=v < modifier_letter_apostrophe [VowelL]=v
[VowelPalCCap]=v < modifier_letter_apostrophe [VowelLCap]=v

;pass(Unicode)

;UniClass [ConsPREiL] = ('d' 'n' 't' 'D' 'N' 'T') ; d n t
```

```
;cyrillic_small_letter_i  <> latin_small_letter_i / ^[ConsPREiL] _ ; 'и'   > ''
;cyrillic_capital_letter_i  <> latin_capital_letter_i  / ^[ConsPREiL] _ ; 'И'   > ''


pass(Unicode)

; long vowels

'i' combining_macron <> 'ii'
'e' combining_macron <> 'ee'
'a' combining_macron <> 'aa'
'o' combining_macron <> 'oo'
'u' combining_macron <> 'uu'
'I' combining_macron <> 'Ii'
'E' combining_macron <> 'Ee'
'A' combining_macron <> 'Aa'
'O' combining_macron <> 'Oo'
'U' combining_macron <> 'Uu'


pass(Unicode)

UniClass [VowelL] = ('i' 'e' 'a' 'o' 'u')

; all other symbols

cyrillic_small_letter_pe  <> latin_small_letter_p ; 'п'   > 'p'
cyrillic_small_letter_be  <> latin_small_letter_b ; 'б'   > 'b'
cyrillic_small_letter_ve  <> latin_small_letter_w ; 'в'   > 'w'
cyrillic_small_letter_em  <> latin_small_letter_m ; 'м'   > 'm'
cyrillic_small_letter_el  <> latin_small_letter_l ; 'л'   > 'l'
cyrillic_small_letter_er  <> latin_small_letter_r ; 'р'   > 'r'
cyrillic_small_letter_te  <> latin_small_letter_t ; 'т'   > 't'
cyrillic_small_letter_de  <> latin_small_letter_d ; 'д'   > 'd'
cyrillic_small_letter_es  <> latin_small_letter_s ; 'с'   > 's'
cyrillic_small_letter_en  <> latin_small_letter_n ; 'н'   > 'n'
cyrillic_small_letter_tse  <> latin_small_letter_c ; 'ц'   > ''
cyrillic_small_letter_che cyrillic_small_letter_soft_sign <> latin_small_letter_c_with_caron ;
'чь'   > 'č'
cyrillic_small_letter_che  <> latin_small_letter_c_with_caron ; 'ч'   > 'č'
cyrillic_small_letter_short_i  > latin_small_letter_j ; 'й'   > 'j'
cyrillic_small_letter_ka  <> latin_small_letter_k ; 'к'   > 'k'
cyrillic_small_letter_ghe <> latin_small_letter_g ; 'г'   > 'g'
cyrillic_small_letter_en_with_hook  > latin_small_letter_eng ; 'ң'   > 'ŋ'
cyrillic_small_letter_en_with_descender  <> latin_small_letter_eng ; 'ң'   > 'ŋ'
cyrillic_small_letter_ha  <> latin_small_letter_x ; 'х'   > 'x'
cyrillic_small_letter_soft_sign <> modifier_letter_apostrophe / _ ^[VowelL] ; 'ь'   > '’'

cyrillic_capital_letter_pe  <> latin_capital_letter_p ; 'п'   > 'p'
cyrillic_capital_letter_be  <> latin_capital_letter_b ; 'б'   > 'b'
cyrillic_capital_letter_ve  <> latin_capital_letter_w ; 'в'   > 'w'
cyrillic_capital_letter_em  <> latin_capital_letter_m ; 'м'   > 'm'
cyrillic_capital_letter_el  <> latin_capital_letter_l ; 'л'   > 'l'
cyrillic_capital_letter_er  <> latin_capital_letter_r ; 'р'   > 'r'
cyrillic_capital_letter_te  <> latin_capital_letter_t ; 'т'   > 't'
cyrillic_capital_letter_de  <> latin_capital_letter_d ; 'д'   > 'd'
cyrillic_capital_letter_es  <> latin_capital_letter_s ; 'с'   > 's'
cyrillic_capital_letter_en  <> latin_capital_letter_n ; 'н'   > 'n'
cyrillic_capital_letter_tse  <> latin_capital_letter_c ; 'Ц'   > ''
cyrillic_capital_letter_che  cyrillic_capital_letter_soft_sign <>
latin_capital_letter_c_with_caron ; 'чь'   > 'č'
cyrillic_capital_letter_che  <> latin_capital_letter_c_with_caron ; 'ч'   > 'č'
cyrillic_capital_letter_short_i  > latin_capital_letter_j ; 'й'   > 'j'
cyrillic_capital_letter_ka  <> latin_capital_letter_k ; 'к'   > 'k'
cyrillic_capital_letter_ghe  <> latin_capital_letter_g ; 'г'   > 'g'
cyrillic_capital_letter_en_with_hook  > latin_capital_letter_eng ; 'ң'   > 'ŋ'
cyrillic_capital_letter_en_with_descender  <> latin_capital_letter_eng ; 'ң'   > 'ŋ'
cyrillic_capital_letter_ha  <> latin_capital_letter_x ; 'х'   > 'x'
cyrillic_capital_letter_soft_sign > modifier_letter_apostrophe; 'ь'   > '’'

; only in Russian loans

cyrillic_small_letter_ef  <> latin_small_letter_f ; 'ф'   > ''
```

```
cyrillic_capital_letter_ef  <> latin_capital_letter_f ; 'Ф'   > ''
cyrillic_small_letter_ze  <> latin_small_letter_z ; 'з'   > ''
cyrillic_capital_letter_ze  <> latin_capital_letter_z ; 'З'   > ''
cyrillic_small_letter_sha  <> latin_small_letter_s_with_caron ; 'ш'   > ''
cyrillic_capital_letter_sha  <> latin_capital_letter_s_with_caron ; 'Ш'   > ''
cyrillic_small_letter_zhe  <> latin_small_letter_z_with_caron ; 'ж'   > ''
cyrillic_capital_letter_zhe  <> latin_capital_letter_z_with_caron ; 'Ж'   > ''
cyrillic_small_letter_shcha  <> latin_small_letter_s_with_caron latin_small_letter_c_with_caron ;
'щ'   > ''
cyrillic_capital_letter_shcha  <> latin_capital_letter_s_with_caron
latin_small_letter_c_with_caron ; 'Щ'   > ''
```

Table 25. Transliteration rules for Logany & Logany 2016

```
pass(Unicode)

UniClass [VowelNonPalC] =     ('ы' 'э' 'а' 'о' 'у') ; 'ы''э''а''о''у'
UniClass [VowelL] =          ('i' 'e' 'a' 'o' 'u') ; 'i''e''a''o''u'
UniClass [VowelNonPalCCap] = ('Ы' 'Э' 'А' 'О' 'У') ; 'Ы''Э''А''О''У'
UniClass [VowelLCap] =         ('I' 'E' 'A' 'O' 'U') ; 'I''E''A''O''U'

; glottal stop

'"' <> latin_letter_glottal_stop
apostrophe <> latin_letter_glottal_stop ;'U+0027' <> latin_letter_glottal_stop

; diphthong

'э(' > 'ä'
'эˇ' > 'ä'

; plain vowels reverse direction

[VowelNonPalC] < [VowelL]
[VowelNonPalCCap] < [VowelLCap]

pass(Unicode)

; и intervocalically

UniClass [VowelC] = ('ы' 'э' 'а' 'о' 'у' 'и' 'е' 'я' 'ё' 'ю' 'ä')

[VowelC] cyrillic_small_letter_i <> [VowelC] latin_small_letter_j latin_small_letter_i


pass(Unicode)

; vowels after D, N (palat.) and other vowels

UniClass [ConsPalC] = ('п' 'в' 'м' 'т' 'д' 'н' 'л' 'ӆ' 'к' 'х' 'ӈ') ;
UniClass [ConsPalCCap] = ('П' 'В' 'М' 'Т' 'Д' 'Н' 'Л' 'Ӆ' 'К' 'Х' 'Ӈ') ;
UniClass [ConsPalL] = ('p' 'w' 'm' 't' 'd' 'n' 'l' 'ɬ' 'k' 'x' 'ŋ') ;
UniClass [ConsPalLCap] = ('P' 'W' 'M' 'T' 'D' 'N' 'L' 'Ɬ' 'K' 'X' 'Ŋ') ;

UniClass [ConsNonPalC] = ('ч''ш''с') ; ч ш
UniClass [ConsNonPalCCap] = ('Ч''Ш''С') ; Ч Ш
UniClass [ConsNonPalL] = ('č''š''s') ; č š
UniClass [ConsNonPalLCap] = ('Č''Š''S') ; Č Š


UniClass [VowelPalC] =        ('и' 'е' 'я' 'ё' 'ю') ;
UniClass [VowelNonPalC] =     ('ы' 'э' 'а' 'о' 'у') ;
UniClass [VowelL] =          ('i' 'e' 'a' 'o' 'u') ;
UniClass [VowelPalXiC] =      ('e' 'я' 'ё' 'ю') ;
UniClass [VowelXiL] =  ('e' 'a' 'o' 'u') ;

UniClass [VowelPalCCap] =       ('И' 'Е' 'Я' 'Ё' 'Ю') ;
UniClass [VowelNonPalCCap] = ('Ы' 'Э' 'А' 'О' 'У') ;
UniClass [VowelLCap] =          ('I' 'E' 'A' 'O' 'U') ;
UniClass [VowelPalXiCCap] =  ('E' 'Я' 'Ё' 'Ю') ;
UniClass [VowelXiLCap] =        ('E' 'A' 'O' 'U') ;


[ConsPalC]=c [VowelPalC]=v > [ConsPalL]=c modifier_letter_apostrophe [VowelL]=v
[ConsNonPalC]=c [VowelPalC]=v > [ConsNonPalL]=c [VowelL]=v
[VowelPalC]=v <> latin_small_letter_j [VowelL]=v
[VowelNonPalC] > [VowelL]

[ConsPalCCap]=c [VowelPalC]=v > [ConsPalLCap]=c modifier_letter_apostrophe [VowelL]=v
[ConsNonPalCCap]=c [VowelPalC]=v > [ConsNonPalLCap]=c [VowelL]=v
[VowelPalCCap]=v <> latin_capital_letter_j [VowelL]=v
[VowelNonPalCCap] > [VowelLCap]

[VowelPalC]=v < modifier_letter_apostrophe [VowelL]=v
[VowelPalCCap]=v < modifier_letter_apostrophe [VowelLCap]=v
```

```
pass(Unicode)

; long vowels

'i' combining_macron <> 'ii'
'e' combining_macron <> 'ee'
'a' combining_macron <> 'aa'
'o' combining_macron <> 'oo'
'u' combining_macron <> 'uu'
'ä' combining_macron <> 'ää'
'I' combining_macron <> 'Ii'
'E' combining_macron <> 'Ee'
'A' combining_macron <> 'Aa'
'O' combining_macron <> 'Oo'
'U' combining_macron <> 'Uu'
'Ä' combining_macron <> 'Ää'


pass(Unicode)

UniClass [VowelL] = ('i' 'e' 'a' 'o' 'u' 'ä')

; all other symbols

cyrillic_small_letter_pe  <> latin_small_letter_p ; 'п'   > 'p'
cyrillic_small_letter_ve  <> latin_small_letter_w ; 'в'   > 'w'
cyrillic_small_letter_em  <> latin_small_letter_m ; 'м'   > 'm'
cyrillic_small_letter_te  <> latin_small_letter_t ; 'т'   > 't'
cyrillic_small_letter_de  <> latin_small_letter_d ; 'д'   > 'd'
cyrillic_small_letter_es  <> latin_small_letter_s ; 'с'   > 's'
cyrillic_small_letter_en  <> latin_small_letter_n ; 'н'   > 'n'
cyrillic_small_letter_el  <> latin_small_letter_l ; 'л'   > 'l'
cyrillic_small_letter_el_with_hook  <> latin_small_letter_l_with_belt ; 'ԯ'   > 'ɬ'
cyrillic_small_letter_che cyrillic_small_letter_soft_sign <> latin_small_letter_c_with_caron ;
'чь'   > 'č'
cyrillic_small_letter_che  <> latin_small_letter_c_with_caron ; 'ч'   > 'č'
cyrillic_small_letter_sha  <> latin_small_letter_s_with_caron ; 'ш'   > ''
cyrillic_small_letter_short_i  > latin_small_letter_j ; 'й'   > 'j'
cyrillic_small_letter_ka  <> latin_small_letter_k ; 'к'   > 'k'
cyrillic_small_letter_en_with_hook  > latin_small_letter_eng ; 'ԋ'   > 'ŋ'
cyrillic_small_letter_en_with_descender  <> latin_small_letter_eng ; 'ң'   > 'ŋ'
cyrillic_small_letter_ha  <> latin_small_letter_x ; 'х'   > 'x'
cyrillic_small_letter_soft_sign <> modifier_letter_apostrophe / _ ^[VowelL] ; 'ь'   > '’'

cyrillic_capital_letter_pe  <> latin_capital_letter_p ; 'п'   > 'p'
cyrillic_capital_letter_ve  <> latin_capital_letter_w ; 'в'   > 'w'
cyrillic_capital_letter_em  <> latin_capital_letter_m ; 'м'   > 'm'
cyrillic_capital_letter_te  <> latin_capital_letter_t ; 'т'   > 't'
cyrillic_capital_letter_de  <> latin_capital_letter_d ; 'д'   > 'd'
cyrillic_capital_letter_es  <> latin_capital_letter_s ; 'с'   > 's'
cyrillic_capital_letter_en  <> latin_capital_letter_n ; 'н'   > 'n'
cyrillic_capital_letter_el  <> latin_capital_letter_l ; 'л'   > 'l'
cyrillic_capital_letter_el_with_hook  <> latin_capital_letter_l_with_belt ; 'ԯ'   > 'ɬ'
cyrillic_capital_letter_che  cyrillic_capital_letter_soft_sign <>
latin_capital_letter_c_with_caron ; 'чь'   > 'č'
cyrillic_capital_letter_che  <> latin_capital_letter_c_with_caron ; 'ч'   > 'č'
cyrillic_capital_letter_sha  <> latin_capital_letter_s_with_caron ; 'Ш'   > ''
cyrillic_capital_letter_short_i  > latin_capital_letter_j ; 'й'   > 'j'
cyrillic_capital_letter_ka  <> latin_capital_letter_k ; 'к'   > 'k'
cyrillic_capital_letter_en_with_hook  > latin_capital_letter_eng ; 'ԋ'   > 'ŋ'
cyrillic_capital_letter_en_with_descender  <> latin_capital_letter_eng ; 'ԋ'   > 'ŋ'
cyrillic_capital_letter_ha  <> latin_capital_letter_x ; 'х'   > 'x'
cyrillic_capital_letter_soft_sign > modifier_letter_apostrophe; 'Ь'   > '’'

; only in Russian loans

cyrillic_small_letter_be  <> latin_small_letter_b ; 'б'   > 'b'
cyrillic_capital_letter_be  <> latin_capital_letter_b ; 'б'   > 'b'
cyrillic_small_letter_ghe <> latin_small_letter_g ; 'г'   > 'g'
cyrillic_capital_letter_ghe  <> latin_capital_letter_g ; 'г'   > 'g'
cyrillic_small_letter_ef  <> latin_small_letter_f ; 'ф'   > ''
```

```
cyrillic_capital_letter_ef  <> latin_capital_letter_f ; 'Ф'   > ''
cyrillic_small_letter_ze  <> latin_small_letter_z ; 'з'   > ''
cyrillic_capital_letter_ze  <> latin_capital_letter_z ; 'З'   > ''
cyrillic_small_letter_er  <> latin_small_letter_r ; 'р'   > 'r'
cyrillic_capital_letter_er  <> latin_capital_letter_r ; 'р'   > 'r'
cyrillic_small_letter_zhe  <> latin_small_letter_z_with_caron ; 'ж'   > ''
cyrillic_capital_letter_zhe  <> latin_capital_letter_z_with_caron ; 'Ж'   > ''
cyrillic_small_letter_tse  <> latin_small_letter_c ; 'ц'   > ''
cyrillic_capital_letter_tse  <> latin_capital_letter_c ; 'Ц'   > ''
cyrillic_small_letter_shcha  <> latin_small_letter_s_with_caron latin_small_letter_c_with_caron ;
'щ'   > ''
cyrillic_capital_letter_shcha  <> latin_capital_letter_s_with_caron
latin_small_letter_c_with_caron ; 'Щ'   > ''
```

Table 26. Transliteration rules for Turutina 2003

```
pass(Unicode)

UniClass [VowelNonPalC] =     ('ы' 'э' 'а' 'о' 'у') ; 'ы''э''а''о''у'
UniClass [VowelL] =          ('і' 'е' 'а' 'о' 'u') ; 'і''е''а''о''u'
UniClass [VowelNonPalCCap] = ('Ы' 'Э' 'А' 'О' 'У') ; 'Ы''Э''А''О''У'
UniClass [VowelLCap] =         ('I' 'E' 'A' 'O' 'U') ; 'I''E''A''O''U'

; glottal stop

'’’' <> latin_letter_glottal_stop
'’' > latin_letter_glottal_stop

; diphthong

'э' combining_breve <> 'ä'
'э□' > 'ä' ; '□' U+F50F, looks like 'э_'
'Э' combining_breve <> 'Ä'
'Э□' > 'Ä' ; '□' U+F50F, looks like 'э_'

; plain vowels reverse direction

[VowelNonPalC] < [VowelL]
[VowelNonPalCCap] < [VowelLCap]

pass(Unicode)

; и intervocalically

UniClass [VowelC] = ('ы' 'э' 'а' 'о' 'у' 'и' 'е' 'я' 'ё' 'ю' 'ä' 'Ы' 'Э' 'А' 'О' 'У' 'И' 'Е' 'Я'
'Ё' 'Ю' 'Ä')

[VowelC] cyrillic_small_letter_i <> [VowelC] latin_small_letter_j latin_small_letter_i


pass(Unicode)

; vowels after D, N (palat.) and other vowels

UniClass [ConsPalC] = ('п' 'в' 'м' 'т' 'д' 'н' 'л' 'ԯ' 'к' 'х' 'ҕ') ;
UniClass [ConsPalCCap] = ('П' 'В' 'М' 'Т' 'Д' 'Н' 'Л' 'Ԯ' 'К' 'Х' 'Ҕ') ;
UniClass [ConsPalL] = ('p' 'w' 'm' 't' 'd' 'n' 'l' 'ɨ' 'k' 'x' 'ŋ') ;
UniClass [ConsPalLCap] = ('P' 'W' 'M' 'T' 'D' 'N' 'L' 'Ɨ' 'K' 'X' 'Ŋ') ;

UniClass [ConsNonPalC] = ('ч''ш''с') ; ч ш
UniClass [ConsNonPalCCap] = ('Ч''Ш''С') ; Ч Ш
UniClass [ConsNonPalL] = ('č''š''s') ; č š
UniClass [ConsNonPalLCap] = ('Č''Š''S') ; Č Š


UniClass [VowelPalC] =        ('и' 'е' 'я' 'ё' 'ю') ;
UniClass [VowelNonPalC] =     ('ы' 'э' 'а' 'о' 'у') ;
UniClass [VowelL] =          ('і' 'е' 'а' 'о' 'u') ;
UniClass [VowelPalXiC] =      ('е' 'я' 'ё' 'ю') ;
UniClass [VowelXiL] = ('e' 'a' 'o' 'u') ;

UniClass [VowelPalCCap] =       ('И' 'Е' 'Я' 'Ё' 'Ю') ;
UniClass [VowelNonPalCCap] = ('Ы' 'Э' 'А' 'О' 'У') ;
UniClass [VowelLCap] =          ('I' 'E' 'A' 'O' 'U') ;
UniClass [VowelPalXiCCap] =  ('Е' 'Я' 'Ё' 'Ю') ;
UniClass [VowelXiLCap] =        ('E' 'A' 'O' 'U') ;


[ConsPalC]=c [VowelPalC]=v > [ConsPalL]=c modifier_letter_apostrophe [VowelL]=v
[ConsNonPalC]=c [VowelPalC]=v > [ConsNonPalL]=c [VowelL]=v
[VowelPalC]=v <> latin_small_letter_j [VowelL]=v
[VowelNonPalC] > [VowelL]

[ConsPalCCap]=c [VowelPalC]=v > [ConsPalLCap]=c modifier_letter_apostrophe [VowelL]=v
[ConsNonPalCCap]=c [VowelPalC]=v > [ConsNonPalLCap]=c [VowelL]=v
[VowelPalCCap]=v <> latin_capital_letter_j [VowelL]=v
```

```
[VowelNonPalCCap] > [VowelLCap]

[VowelPalC]=v < modifier_letter_apostrophe [VowelL]=v
[VowelPalCCap]=v < modifier_letter_apostrophe [VowelLCap]=v


pass(Unicode)

; long vowels

'i' combining_macron <> 'ii'
'e' combining_macron <> 'ee'
'a' combining_macron <> 'aa'
'o' combining_macron <> 'oo'
'u' combining_macron <> 'uu'
'ä' combining_macron <> 'ää'
'I' combining_macron <> 'Ii'
'E' combining_macron <> 'Ee'
'A' combining_macron <> 'Aa'
'O' combining_macron <> 'Oo'
'U' combining_macron <> 'Uu'
'Ä' combining_macron <> 'Ää'


pass(Unicode)

UniClass [VowelL] = ('i' 'e' 'a' 'o' 'u' 'ä')

; all other symbols

cyrillic_small_letter_pe  <> latin_small_letter_p ; 'п'   > 'p'
cyrillic_small_letter_ve  <> latin_small_letter_w ; 'в'   > 'w'
cyrillic_small_letter_em  <> latin_small_letter_m ; 'м'   > 'm'
cyrillic_small_letter_te  <> latin_small_letter_t ; 'т'   > 't'
cyrillic_small_letter_de  <> latin_small_letter_d ; 'д'   > 'd'
cyrillic_small_letter_es  <> latin_small_letter_s ; 'с'   > 's'
cyrillic_small_letter_en  <> latin_small_letter_n ; 'н'   > 'n'
cyrillic_small_letter_el  <> latin_small_letter_l ; 'л'   > 'l'
cyrillic_small_letter_el_with_hook  <> latin_small_letter_l_with_belt ; 'ԯ'   > 'ɬ'
cyrillic_small_letter_che cyrillic_small_letter_soft_sign <> latin_small_letter_c_with_caron ;
'чь'   > 'č'
cyrillic_small_letter_che  <> latin_small_letter_c_with_caron ; 'ч'   > 'č'
cyrillic_small_letter_sha  <> latin_small_letter_s_with_caron ; 'ш'   > ''
cyrillic_small_letter_short_i  > latin_small_letter_j ; 'й'   > 'j'
cyrillic_small_letter_ka  <> latin_small_letter_k ; 'к'   > 'k'
cyrillic_small_letter_en_with_hook  > latin_small_letter_eng ; 'ԩ'   > 'ŋ'
cyrillic_small_letter_en_with_descender  <> latin_small_letter_eng ; 'ԣ'   > 'ŋ'
cyrillic_small_letter_ha  <> latin_small_letter_x ; 'х'   > 'x'
cyrillic_small_letter_soft_sign <> modifier_letter_apostrophe / _ ^[VowelL] ; 'ь'   > '’'

cyrillic_capital_letter_pe  <> latin_capital_letter_p ; 'П'   > 'p'
cyrillic_capital_letter_ve  <> latin_capital_letter_w ; 'В'   > 'w'
cyrillic_capital_letter_em  <> latin_capital_letter_m ; 'М'   > 'm'
cyrillic_capital_letter_te  <> latin_capital_letter_t ; 'Т'   > 't'
cyrillic_capital_letter_de  <> latin_capital_letter_d ; 'Д'   > 'd'
cyrillic_capital_letter_es  <> latin_capital_letter_s ; 'С'   > 's'
cyrillic_capital_letter_en  <> latin_capital_letter_n ; 'Н'   > 'n'
cyrillic_capital_letter_el  <> latin_capital_letter_l ; 'Л'   > 'l'
cyrillic_capital_letter_el_with_hook  <> latin_capital_letter_l_with_belt ; 'Ԯ'   > 'ɬ'
cyrillic_capital_letter_che  cyrillic_capital_letter_soft_sign <>
latin_capital_letter_c_with_caron ; 'чь'   > 'č'
cyrillic_capital_letter_che  <> latin_capital_letter_c_with_caron ; 'Ч'   > 'č'
cyrillic_capital_letter_sha  <> latin_capital_letter_s_with_caron ; 'Ш'   > ''
cyrillic_capital_letter_short_i  > latin_capital_letter_j ; 'Й'   > 'j'
cyrillic_capital_letter_ka  <> latin_capital_letter_k ; 'К'   > 'k'
cyrillic_capital_letter_en_with_hook  > latin_capital_letter_eng ; 'Ԩ'   > 'ŋ'
cyrillic_capital_letter_en_with_descender  <> latin_capital_letter_eng ; 'Ԣ'   > 'ŋ'
cyrillic_capital_letter_ha  <> latin_capital_letter_x ; 'Х'   > 'x'
cyrillic_capital_letter_soft_sign > modifier_letter_apostrophe; 'Ь'   > '’'

; only in Russian loans

cyrillic_small_letter_be  <> latin_small_letter_b ; 'б'   > 'b'
```

```
cyrillic_capital_letter_be  <> latin_capital_letter_b ; 'б'   > 'b'
cyrillic_small_letter_ghe <> latin_small_letter_g ; 'г'   > 'g'
cyrillic_capital_letter_ghe  <> latin_capital_letter_g ; 'г'   > 'g'
cyrillic_small_letter_ef  <> latin_small_letter_f ; 'ф'   > ''
cyrillic_capital_letter_ef  <> latin_capital_letter_f ; 'Ф'   > ''
cyrillic_small_letter_ze  <> latin_small_letter_z ; 'з'   > ''
cyrillic_capital_letter_ze  <> latin_capital_letter_z ; 'З'   > ''
cyrillic_small_letter_er  <> latin_small_letter_r ; 'р'   > 'r'
cyrillic_capital_letter_er  <> latin_capital_letter_r ; 'р'   > 'r'
cyrillic_small_letter_zhe  <> latin_small_letter_z_with_caron ; 'ж'   > ''
cyrillic_capital_letter_zhe  <> latin_capital_letter_z_with_caron ; 'Ж'   > ''
cyrillic_small_letter_tse  <> latin_small_letter_c ; 'ц'   > ''
cyrillic_capital_letter_tse  <> latin_capital_letter_c ; 'Ц'   > ''
cyrillic_small_letter_shcha  <> latin_small_letter_s_with_caron latin_small_letter_c_with_caron ;
'щ'   > ''
cyrillic_capital_letter_shcha  <> latin_capital_letter_s_with_caron
latin_small_letter_c_with_caron ; 'Щ'   > ''
```

Table 27. Transliteration for Yangasova 2018

```
pass(Unicode)

UniClass [VowelNonPalC] =     ('ы' 'э' 'а' 'о' 'у') ; 'ы''э''а''о''у'
UniClass [VowelL] =          ('i' 'e' 'a' 'o' 'u') ; 'i''e''a''o''u'
UniClass [VowelNonPalCCap] = ('Ы' 'Э' 'А' 'О' 'У') ; 'Ы''Э''А''О''У'
UniClass [VowelLCap] =         ('I' 'E' 'A' 'O' 'U') ; 'I''E''A''O''U'

; glottal stop

'”' <> latin_letter_glottal_stop ; U+201D (in Burkova-TB)
'’' <> latin_letter_glottal_stop ; U+2019 (in Burkova-TB)

; cyrillic vowels with macron

cyrillic_small_letter_i_with_macron <> cyrillic_small_letter_i combining_macron ;й
cyrillic_small_letter_u_with_macron <> cyrillic_small_letter_u combining_macron ;ў
cyrillic_capital_letter_i_with_macron <> cyrillic_capital_letter_i combining_macron ;Й
cyrillic_capital_letter_u_with_macron <> cyrillic_capital_letter_u combining_macron ;Ў
cyrillic_small_letter_short_u <> cyrillic_small_letter_u combining_breve ;ў
cyrillic_capital_letter_short_u <> cyrillic_capital_letter_u combining_breve ;Ў

; plain vowels reverse direction

[VowelNonPalC] < [VowelL]
[VowelNonPalCCap] < [VowelLCap]

pass(Unicode)

; и intervocalically

UniClass [VowelC] = ('ы' 'э' 'а' 'о' 'у' 'и' 'е' 'я' 'ё' 'ю' 'ä' 'Ы' 'Э' 'А' 'О' 'У' 'И' 'Е' 'Я'
'Ё' 'Ю' 'Ä')

[VowelC] cyrillic_small_letter_i <> [VowelC] latin_small_letter_j latin_small_letter_i


pass(Unicode)

; vowels after D, N (palat.) and other vowels

UniClass [ConsPalC] = ('п' 'б' 'в' 'м' 'т' 'д' 'н' 'л' 'р' 'ц' 'с' 'з' 'к' 'х' 'ң' 'ң' 'ҕ' 'ҕ') ;
'ң' (in Burkova-TB)
UniClass [ConsPalCCap] = ('П' 'Б' 'В' 'М' 'Т' 'Д' 'Н' 'Л' 'Р' 'Ц' 'С' 'З' 'К' 'Х' 'Ң' 'Ң' 'Ҕ'
'Ҕ') ; 'Ң' (in Burkova-TB)
UniClass [ConsPalL] = ('p' 'b' 'w' 'm' 't' 'd' 'n' 'l' 'r' 'c' 's' 'z' 'k' 'x' 'ŋ' 'ŋ' 'ŋ''ŋ')
UniClass [ConsPalLCap] = ('P' 'B' 'W' 'M' 'T' 'D' 'N' 'L' 'R' 'C' 'S' 'Z' 'K' 'X' 'Ŋ' 'Ŋ' 'Ŋ''Ŋ')

UniClass [ConsNonPalC] = ('ч''ш') ; ч ш
UniClass [ConsNonPalCCap] = ('Ч''Ш') ; Ч Ш
UniClass [ConsNonPalL] = ('č''š') ; č š
UniClass [ConsNonPalLCap] = ('Č''Š') ; Č Š


UniClass [VowelPalC] =        ('и' 'е' 'я' 'ё' 'ю') ;
UniClass [VowelNonPalC] =     ('ы' 'э' 'а' 'о' 'у') ;
UniClass [VowelL] =          ('i' 'e' 'a' 'o' 'u') ;
UniClass [VowelPalXiC] =      ('е' 'я' 'ё' 'ю') ;
UniClass [VowelXiL] = ('e' 'a' 'o' 'u') ;

UniClass [VowelPalCCap] =       ('И' 'Е' 'Я' 'Ё' 'Ю') ;
UniClass [VowelNonPalCCap] = ('Ы' 'Э' 'А' 'О' 'У') ;
UniClass [VowelLCap] =          ('I' 'E' 'A' 'O' 'U') ;
UniClass [VowelPalXiCCap] =  ('Е' 'Я' 'Ё' 'Ю') ;
UniClass [VowelXiLCap] =        ('E' 'A' 'O' 'U') ;


[ConsPalC]=c [VowelPalC]=v > [ConsPalL]=c modifier_letter_apostrophe [VowelL]=v
[ConsNonPalC]=c [VowelPalC]=v > [ConsNonPalL]=c [VowelL]=v
[VowelPalC]=v <> latin_small_letter_j [VowelL]=v
[VowelNonPalC] > [VowelL]
```

```
[ConsPalCCap]=c [VowelPalC]=v > [ConsPalLCap]=c modifier_letter_apostrophe [VowelL]=v
[ConsNonPalCCap]=c [VowelPalC]=v > [ConsNonPalLCap]=c [VowelL]=v
[VowelPalCCap]=v <> latin_capital_letter_j [VowelL]=v
[VowelNonPalCCap] > [VowelLCap]

[VowelPalC]=v < modifier_letter_apostrophe [VowelL]=v
[VowelPalCCap]=v < modifier_letter_apostrophe [VowelLCap]=v


pass(Unicode)

; long vowels

'i' combining_macron <> 'ii'
'e' combining_macron <> 'ee'
'a' combining_macron <> 'aa'
'o' combining_macron <> 'oo'
'u' combining_macron <> 'uu'
'ä' combining_macron <> 'ää'
'I' combining_macron <> 'Ii'
'E' combining_macron <> 'Ee'
'A' combining_macron <> 'Aa'
'O' combining_macron <> 'Oo'
'U' combining_macron <> 'Uu'
'Ä' combining_macron <> 'Ää'


pass(Unicode)

UniClass [VowelL] = ('i' 'e' 'a' 'o' 'u' 'ä')

; all other symbols

cyrillic_small_letter_pe  <> latin_small_letter_p ; 'п'   > 'p'
cyrillic_small_letter_ve  <> latin_small_letter_w ; 'в'   > 'w'
cyrillic_small_letter_em  <> latin_small_letter_m ; 'м'   > 'm'
cyrillic_small_letter_te  <> latin_small_letter_t ; 'т'   > 't'
cyrillic_small_letter_de  <> latin_small_letter_d ; 'д'   > 'd'
cyrillic_small_letter_es  <> latin_small_letter_s ; 'с'   > 's'
cyrillic_small_letter_en  <> latin_small_letter_n ; 'н'   > 'n'
cyrillic_small_letter_el  <> latin_small_letter_l ; 'л'   > 'l'
cyrillic_small_letter_tse  <> latin_small_letter_c ; 'ц'   > ''
cyrillic_small_letter_che cyrillic_small_letter_soft_sign <> latin_small_letter_c_with_caron ;
'чь'   > 'č'
cyrillic_small_letter_che  <> latin_small_letter_c_with_caron ; 'ч'   > 'č'
cyrillic_small_letter_sha  <> latin_small_letter_s_with_caron ; 'ш'   > 'š'
cyrillic_small_letter_shcha  <> latin_small_letter_s_with_acute ; 'щ'   > 'ś' in Koshkareva 2005
cyrillic_small_letter_short_i  > latin_small_letter_j ; 'й'   > 'j'
cyrillic_small_letter_ka  <> latin_small_letter_k ; 'к'   > 'k'
cyrillic_small_letter_en_with_tail  > latin_small_letter_eng ; 'ң'   > 'ŋ'  (in Burkova-TB)
cyrillic_small_letter_en_with_descender  > latin_small_letter_eng ; 'ң'   > 'ŋ'  (in Burkova-TB)
cyrillic_small_letter_ha  <> latin_small_letter_x ; 'х'   > 'x'
cyrillic_small_letter_soft_sign <> modifier_letter_apostrophe / _ ^[VowelL] ; 'ь'   > ''
cyrillic_small_letter_hard_sign > ; 'ъ' in Koshkareva 2005

cyrillic_capital_letter_pe  <> latin_capital_letter_p ; 'п'   > 'p'
cyrillic_capital_letter_ve  <> latin_capital_letter_w ; 'в'   > 'w'
cyrillic_capital_letter_em  <> latin_capital_letter_m ; 'м'   > 'm'
cyrillic_capital_letter_te  <> latin_capital_letter_t ; 'т'   > 't'
cyrillic_capital_letter_de  <> latin_capital_letter_d ; 'д'   > 'd'
cyrillic_capital_letter_es  <> latin_capital_letter_s ; 'с'   > 's'
cyrillic_capital_letter_en  <> latin_capital_letter_n ; 'н'   > 'n'
cyrillic_capital_letter_el  <> latin_capital_letter_l ; 'л'   > 'l'
cyrillic_capital_letter_tse  <> latin_capital_letter_c ; 'Ц'   > ''
cyrillic_capital_letter_che  cyrillic_capital_letter_soft_sign <>
latin_capital_letter_c_with_caron ; 'чь'   > 'č'
cyrillic_capital_letter_che  <> latin_capital_letter_c_with_caron ; 'ч'   > 'č'
cyrillic_capital_letter_sha  <> latin_capital_letter_s_with_caron ; 'Ш'   > ''
cyrillic_capital_letter_shcha  <> latin_capital_letter_s_with_acute ; 'Щ'   > 'Ś' in Koshkareva
2005
cyrillic_capital_letter_short_i  > latin_capital_letter_j ; 'й'   > 'j'
cyrillic_capital_letter_ka  <> latin_capital_letter_k ; 'к'   > 'k'
cyrillic_capital_letter_en_with_tail  > latin_capital_letter_eng ; 'ң'   > 'ŋ'  (in Burkova-TB)
```

```
cyrillic_capital_letter_en_with_descender  > latin_capital_letter_eng ; 'Ҥ'   > 'ŋ'  (in Burkova-
TB)
'ҥ' > 'ŋ'
'Ҕ' > 'Ɲ'
cyrillic_capital_letter_ha  <> latin_capital_letter_x ; 'х'   > 'x'
cyrillic_capital_letter_soft_sign > modifier_letter_apostrophe; 'Ь'   > 'ʼ'
cyrillic_capital_letter_hard_sign > ; 'ъ' in Koshkareva 2005

; only in Russian loans

cyrillic_small_letter_be  <> latin_small_letter_b ; 'б'   > 'b'
cyrillic_capital_letter_be  <> latin_capital_letter_b ; 'б'   > 'b'
cyrillic_small_letter_ghe <> latin_small_letter_g ; 'г'   > 'g'
cyrillic_capital_letter_ghe  <> latin_capital_letter_g ; 'г'   > 'g'
cyrillic_small_letter_ef  <> latin_small_letter_f ; 'ф'   > ''
cyrillic_capital_letter_ef  <> latin_capital_letter_f ; 'Ф'   > ''
cyrillic_small_letter_ze  <> latin_small_letter_z ; 'з'   > ''
cyrillic_capital_letter_ze  <> latin_capital_letter_z ; 'З'   > ''
cyrillic_small_letter_er  <> latin_small_letter_r ; 'р'   > 'r'
cyrillic_capital_letter_er  <> latin_capital_letter_r ; 'р'   > 'r'
cyrillic_small_letter_zhe  <> latin_small_letter_z_with_caron ; 'ж'   > ''
cyrillic_capital_letter_zhe  <> latin_capital_letter_z_with_caron ; 'Ж'   > ''


pass(Unicode)

'«' > '"'
'»' > '"'
```

Appendix A2. Transliteration in metadata

All personal names and most placenames in respective metadata fields are transliterated from Cyrillic into Latin alphabet following the transliteration standard GOST 7.79 System B (published as GOST 2001).

Elsewhere, e.g. in text titles, English glosses (**ge** tier) and free translations (**fe** tier), English-style romanization is used.

Table 28. INEL transliteration of Russian letters (GOST 7.79 System B)

Cyrillic		**Roman**		**Notes**
А	а	A	a	
Б	б	B	b	
В	в	V	v	
Г	г	G	g	
Д	д	D	d	
Е	е	E	e	
Ё	ё	Yo	yo	
Ж	ж	Zh	zh	
З	з	Z	z	
И	и	I	i	
Й	й	J	j	
К	к	K	k	
Л	л	L	l	
М	м	M	m	
Н	н	N	n	
О	о	O	o	
П	п	P	p	
Р	р	R	r	
С	с	S	s	
Т	т	T	t	
У	у	U	u	
Ф	ф	F	f	
Х	х	X	x	
Ц	ц	C	c	INEL uses c everywhere (cz is recommended in GOST when not before i, e, y, j)
Ч	ч	Ch	ch	
Ш	ш	Sh	sh	
Щ	щ	Shh	shh	
Ъ	ъ	``	``	double grave accent
Ы	ы	Y`	y`	
Ь	ь	`	`	grave accent
Э	э	E`	e`	
Ю	ю	Yu	yu	
Я	я	Ya	ya	

Table 29. INEL Russian to English romanization

Cyrillic		English		Note
А	а	A	a	
Б	б	B	b	
В	в	V	v	
Г	г	G	g	
Д	д	D	d	
Е	е	Ye / e	ye / e	*ye* word-initially, after a vowel or ъ; *e* elsewhere E.g. *Енисей > Yenisey, Большое > Bolshoye*
Ё	ё	Yo	yo	
Ж	ж	Zh	zh	
З	з	Z	z	
И	и	I	i	
	-ий		-y / -iy	*y* in Russian endings, e.g. *Самарский > Samarsky* *iy* in Enets or Nenets names, e.g. *Волебий > Volebiy*
Й	й	Y	y	
К	к	K	k	
Л	л	L	l	
М	м	M	m	
Н	н	N	n	
О	о	O	o	
П	п	P	p	
Р	р	R	r	
С	с	S	s	
Т	т	T	t	
У	у	U	u	
Ф	ф	F	f	
Х	х	Kh	kh	
Ц	ц	Ts	ts	
Ч	ч	Ch	ch	
Ш	ш	Sh	sh	
Щ	щ	Shch	shch	
–	ъ	–	omitted	
Ы	ы	Y	y	
–	-ый	–	-y / -iy	*y* in Russian endings, e.g. *Малый > Maly* *iy* in Enets or Nenets names, e.g. *Ласый > Lasiy*
–	ь	–	omitted	E.g. *Татьяна > Tatyana, Ванька > Vanka*
–	-ье	–	-ye	E.g. *Афанасьевич > Afanasyevich*
–	-ьо	–	-yo	E.g. *Бульон > Bulyon*
Э	э	E	e	
Ю	ю	Yu	yu	
Я	я	Ya	ya	

Appendix A3. Nenets speakers

Appendix A43 contains the full lists of speakers who contributed to the corpus as storytellers and/or who provided oral transcription of texts with their codes, we gratefully acknowledge everyone's contribution and realize that this corpus could never be possible, unless as a result of a joint effort of all the people listed here.

In this Appendix, transliteration as in the metadata is used (see 2.5.6).

Table 30. Forest Nenets speakers

Name	**Also known as**	**Main residence place**	**Code**
Ajvaseda, Ajty` Kol`chevna	Najty`	Var`yogan	AAK
Ajvaseda, Aleksandr Teklevich		Var`yogan	AAT
Ajvaseda, Nyonyu Kol`chevna		Var`yogan	ANK
Ajvaseda, Pavel Yanchevich		...	APYa
Ajvaseda, Vaxalyuma Yankleevich		Var`yogan	AVYa
Ajvaseda, Vladimir Talyovich		Var`yogan	AVT
Ajvaseda, Yurij Ky`levich		Var`yogan	AYuK
Ajvasedo, Lyubov` Y`nevna		Voento	ALY
Iusi, Lyudmila Aulevna		Var`yogan	ILA
Kary`mova, Liliya Alevna		Var`yogan	KLA
Kazamkina, Tateni Sobolevna		Var`yogan	KTS
Logany`, Aleksandr Chapixovich	Tatva	Porsavar	LACh
Pyak, Irina Kajlevna		...	PIK
Pyak, Lemya Kamchatovich		village Numto	PLK
Pyak, Semyon My`tovich	Shunya	village Numto	PSM
Pyak, Vasilij Lemevich	Yun`toma	Pichmeej, Oxlor Bay, Lake Numto	PVL
Pyak, Vladimir My`tovich	Shalyupi	camp on the Sorum Kasam river	PVM
Randimova, Evdokiya		...	RE
Solina, Anna Osevna	Uxuli	Shapsha	SAO
Turutina, Polina Gilevna		Tarko-Sale	TPG
Ve`llo, Anely`		Xarampur	VA
Ve`llo, Roza		Xarampur	VR

Table 31. Tundra Nenets speakers

Name	**Main residence place**	**Code**
Anufrieva, Ekaterina Petrovna	Nady`m	AEP
Ardeeva, E.	...	AE
Bobrikov, Vasilij Semyonovich	Kanin tundra	BVS
Bobrikova, Evdokiya Yakovlevna	Chizha	BEYa
Bobrikova, Parakov`ya Konstantinovna	Chizha	BPK
Kanyukova, Anna Pavlovna	Nar`yan-Mar	KAP
Ladukaj, Aleksej Ivanovich	Yar-Sale	LAI
Laty`sheva, Anna Fyodorovna	...	LAF
Laty`sheva, Elena Artem`evna	...	LEA
Laty`sheva, Tat`yana	...	LT
Ledkov, Ivan Patrovich	Chyornaya	LIP
Ledkova, Irin`ya Stepanovna	Kanin tundra	LIS
Ledkova, Mariya Mixajlovna	Chyornaya	LMM
Ly`rmina, Ekaterina Petrovna	Voroncovo	LEP
Mar``ik, Mariya Nikolaevna	Selyanskie peski	MMN
Nyurov, Artemij Ivanovich	tundra	NAI
Nyurova, Oktyabrina Ivanovna	...	NOI
Salender, Elena	...	SE

Name	Main residence place	Code
Sulen`ev, Nikolaj Pavlovich	Chizha	SNP
Tajbarej, Evgeniya Grigor`evna	Chyornaya	TEG
Taleev, Efim Grigor`evich	Malozemel`skaya tundra	TEfG
Ve`lla, Valentina Matkalievna	Nori	VVM
Vengo, Ejko	...	VE
Yande, Evdokiya Andreevna	Tuxard	YaEA
Yande, Naka Nonovich	Tuxard	YaNN
Yande, Nina Mixajlovna	Tuxard	YaNM
Yando, Monguchi Linovich	Voroncovo	YaML
Yaptune`, Ochavka Yujnovich	Nosok	YaOYu
Yaptune`, Svetlana Pelevna	Nosok	YaSP
Yar, Vasilij Dmirievich	Voroncovo	YaVD
Yar, Yavmbya Andreevich	Voroncovo	YaYA

Appendix A4. Morpheme glossing labels (tiers ge, gr) and Tsakorpus grammar tags

The following tables list the glossing labels used in tiers **ge**, **gr** and corresponding grammar tags for use in Tsakorpus online search.

Table 32. List of morpheme glossing labels by category

Gloss	Description	Tsakorpus grammar tags	Tsakorpus category	Comment
Person and number				
1SG	1 person singular	pn1,pnsg	persnum-person, persnum-number	
1DU	1 person dual	pn1,pndu	persnum-person, persnum-number	
1PL	1 person plural	pn1,pnpl	persnum-person, persnum-number	
2SG	2 person singular	pn2,pnsg	persnum-person, persnum-number	
2DU	2 person dual	pn2,pndu	persnum-person, persnum-number	
2PL	2 person plural	pn2,pnpl	persnum-person, persnum-number	
3SG	3 person singular	pn3,pnsg	persnum-person, persnum-number	
3DU	3 person dual	pn3,pndu	persnum-person, persnum-number	
3PL	3 person plural	pn3,pnpl	persnum-person, persnum-number	
Personal pronouns				
	personal pronoun	pers	pers	only as a tag
PRO1SG	personal pronoun, 1 person singular	pers,pn1,pnsg	pers,persnum-person, persnum-number	lexical
PRO1DU	personal pronoun, 1 person dual	pers,pn1,pndu	pers,persnum-person, persnum-number	lexical
PRO1PL	personal pronoun, 1 person plural	pers,pn1,pnpl	pers,persnum-person, persnum-number	lexical
PRO2SG	personal pronoun, 2 person singular	pers,pn2,pnsg	pers,persnum-person, persnum-number	lexical
PRO2DU	personal pronoun, 2 person dual	pers,pn2,pndu	pers,persnum-person, persnum-number	lexical
PRO2PL	personal pronoun, 2 person plural	pers,pn2,pnpl	pers,persnum-person, persnum-number	lexical
PRO3SG	personal pronoun, 3 person singular	pers,pn3,pnsg	pers,persnum-person, persnum-number	lexical
PRO3DU	personal pronoun, 3 person dual	pers,pn3,pndu	pers,persnum-person, persnum-number	lexical
PRO3PL	personal pronoun, 3 person plural	pers,pn3,pnpl	pers,persnum-person, persnum-number	lexical
Nominal categories				
Number				
SG	singular number	sg	n-num	only in combinations
DU	dual number	du	n-num	
PL	plural number	pl	n-num	
Case				
ABL	ablative case	abl	n-case	
ACC	accusative case	acc	n-case	
ESS	essive case	ess	n-case	

Gloss	Description	Tsakorpus grammar tags	Tsakorpus category	Comment
GEN	genitive case	gen	n-case	
LAT	lative case	lat	n-case	
LOC	locative case	loc	n-case	
NOM	nominative case	nom	n-case	only in combinations, e.g. “NOM.3SG”
OBL	oblique case	obl	n-case	only in combinations, e.g. “OBL.1SG”
PROL	prolative case	prol	n-case	
VOC	vocative case	voc	n-case	
Possession				
	possessive	poss	poss	only as a tag
Other nominal categories				
ABL.ADV	ablative case of adverbials	abl,advcase	n-case,n-misc	
DST	destinative	dst	n-misc	
DST.PL	destinative, plural number	dst,pl	n-misc,n-num	
LAT.ADV	lative case of adverbials	lat,advcase	n-case,n-misc	
LOC.ADV	locative case of adverbials	loc,advcase	n-case,n-misc	
Verbal categories				
Conjugation type				
S	subjective conjugation	subjc	persnum-conj	only in combinations
O	objective conjugation	objc	persnum-conj	only in combinations
SG.O	objective conjugation, singular object	objc,objsg	persnum-conj, persnum-num	only in combinations
NSG.O	objective conjugation, nonsingular object	objc,objnsg	persnum-conj, persnum-num	only in combinations
MD	middle conjugation	md	persnum-conj	only in combinations
TAM categories				
APPR	apprehensive	appr	v-tam	
AUD	auditive	aud	v-tam	
	dubitative	dub	v-tam	only as a tag
DUB1	dubitative 1	dub1	v-tam	
DUB2	dubitative 2	dub2	v-tam	
DUR	durative	dur	v-tam	
EVID	evidential	evid	v-tam	
EVID.PRS	evidential, present tense	evid,prs	v-tam,v-tam	
EVID.FUT	evidential, future tense	evid,fut	v-tam,v-tam	
EVID.PST	evidential, past tense	evid,pst	v-tam,v-tam	
FRQ	frequentative	frq	v-tam	
FUT	future tense	fut	v-tam	
HAB	habitual	hab	v-tam	
HORT	hortative	hort	v-tam	
IMP	imperative	imp	v-tam	
INCH	inchoative	inch	v-tam	
INTER	interrogative	inter	v-tam	
IPFV	imperfectivizer	ipfv	v-tam	
JUSS	jussive	juss	v-tam	
MOM	momentaneous	mom	v-tam	
NAR	narrative	nar	v-tam	
NEC	necessitative	nec	v-tam	

Gloss	Description	Tsakorpus grammar tags	Tsakorpus category	Comment
OPT	optative	opt	v-tam	
PROB	probabilitive	prob	v-tam	
PROG	progressive	prog	v-tam	
PROS	prospective	pros	v-tam	
PRS	present tense	prs	v-tam	only in combinations
PST	past tense	pst	v-tam	
SPEC	speculative	spec	v-tam	
SPEC.PRS	speculative, present tense	spec,prs	v-tam,v-tam	
SPEC.PST	speculative, past tense	spec,pst	v-tam,v-tam	
SPEC.FUT	speculative, future tense	spec,fut	v-tam,v-tam	
%SPEC	possibly speculative	uncert,spec	misc,v-tam	
Non-finite forms				
	conditional converb	cond	v-nfin	only as a tag
COND1	conditional converb 1	cond,cond1	v-nfin,v-nfin	
COND2	conditional converb 2	cond,cond2	v-nfin,v-nfin	
CNG	connegative	cng	v-nfin	
CVB	Converb	cnvb	v-nfin	
CVB.IMM	immediate converb	cnvb,imm	v-nfin,v-nfin	
INF	infinitive	inf	v-nfin	
NMLZ	nominalization	nmlz	v-nfin	
PTCP	participle	ptcp	v-nfin	only in combinations
PTCP.PRS	present participle	ptcp,prs	v-nfin,v-tam	
PTCP.PST	past participle	ptcp,pst	v-nfin,v-tam	
PTCP.FUT	future participle	ptcp,fut	v-nfin,v-tam	
PTCP.NEC	necessitative participle	ptcp,nec	v-nfin,v-tam	
PTCP.RES	resultative participle	ptcp.res	v-nfin	
PTCP.ABES	abessive participle	ptcp,abes	v-nfin,v-nfin	
SUP	supine	sup	v-nfin	
Negation				
NEG	negation	neg	neg	lexical
NEG.EMPH	emphatic negative	neg,emph	neg,neg	lexical
NEG.IMP	negative imperative	neg,imp	neg,neg	lexical
Nominal derivations				
	action nominal (any)	act	deriv-n	only as a tag
ACT1	action nominal 1	act1	deriv-n	
ACT2	action nominal 2	act2	deriv-n	
ACT3	action nominal 3	act3	deriv-n	
ACT4	action nominal 4	act4	deriv-n	
AUG	augmentative	aug	deriv-n	
	diminutive (any)	dim	deriv-n	only as a tag
DIM1	diminutive 1	dim,dim1	deriv-n,deriv-n	
DIM2	diminutive 2	dim,dim2	deriv-n,deriv-n	
DIM3	diminutive 3	dim,dim3	deriv-n,deriv-n	
DIM4	diminutive 4	dim,dim4	deriv-n,deriv-n	
DIM5	diminutive 5	dim,dim5	deriv-n,deriv-n	
DIM6	diminutive 6	dim,dim6	deriv-n,deriv-n	
DIM7	diminutive 7	dim,dim7	deriv-n,deriv-n	
DIM8	diminutive 8	dim,dim8	deriv-n,deriv-n	
DYA	dyadic (connective-reciprocal)	dya	deriv-n	
INSTR	instrumental nominal	instr	deriv-n	
LOCN	location nominal	locn	deriv-n	
MOD	modifier	mod	deriv-n	

Gloss	Description	Tsakorpus grammar tags	Tsakorpus category	Comment
PEJ	pejorative	pej	deriv-n	
Verbal derivations				
ATT	attenuative	att	deriv-v	
CAP	captative	cap	deriv-v	
CAUS	causative	caus	deriv-v	
DRV	unspecified derivation	drv	deriv-v	
INT	intensifier	int	deriv-v	
REP	reportative	rep	deriv-v	
%REP	possibly reportative	uncert,rep	misc,deriv-v	
PASS	passive participle	pass	deriv-v	
TR	transitivizer	tr	deriv-v	
Miscellaneous				
ADJZ	adjectivizer	adjz	deriv-misc	
ADVZ	adjverbializer	advz	deriv-misc	
CAR	caritive	car	deriv-misc	
CO	coaffix	co	misc	
DSTR	distributive	dstr	deriv-misc	
EP	epenthetic element	ep	misc	
LIM	limitative	lim	deriv-misc	
LOCZ	locativizer	locz	deriv-misc	
ORD	ordinal numeral	ord	deriv-misc	
POSSV	possessive verbalizer	possv	deriv-misc	
PROP	proprietive	prop	deriv-misc	
RAR	raritive	rar	deriv-misc	
SEL	selective	sel	deriv-misc	
SIM	similative	sim	deriv-misc	
TRL	translative	trl	deriv-misc	
VBLZ	verbalizer	vblz	deriv-misc	
CTR	contrastive	ctr	disc	lexical
EMPH	emphatic marker	emph	disc	lexical
EXCL	exclamative	excl	disc	lexical
INDF	indefinite marker	indf	disc	lexical
INTERJ	unspecified interjection	interj	disc	lexical
PLC	placeholder stem	plc	misc	
TOP	topical marker	top	disc	lexical
%	uncertain categories	uncert	misc	as part of a gloss, e.g. “%DRV”
%%	unknown morph	unkn	misc	

Table 33. Alphabetical list of morpheme glossing labels

Gloss	Description	Tsakorpus grammar tags	Tsakorpus category	Comment
1DU	1 person dual	pn1,pndu	persnum-person, persnum-number	
1PL	1 person plural	pn1,pnpl	persnum-person, persnum-number	
1SG	1 person singular	pn1,pnsg	persnum-person, persnum-number	
2DU	2 person dual	pn2,pndu	persnum-person, persnum-number	
2PL	2 person plural	pn2,pnpl	persnum-person, persnum-number	
2SG	2 person singular	pn2,pnsg	persnum-person, persnum-number	
3DU	3 person dual	pn3,pndu	persnum-person, persnum-number	
3PL	3 person plural	pn3,pnpl	persnum-person, persnum-number	
3SG	3 person singular	pn3,pnsg	persnum-person, persnum-number	
ABL	ablative case	abl	n-case	
ABL.ADV	ablative case of adverbials	abl,advcase	n-case,n-misc	
ACC	accusative case	acc	n-case	
ACT1	action nominal 1	act1	deriv-n	
ACT2	action nominal 2	act2	deriv-n	
ACT3	action nominal 3	act3	deriv-n	
ACT4	action nominal 4	act4	deriv-n	
ADJZ	adjectivizer	adjz	deriv-misc	
ADVZ	adjverbializer	advz	deriv-misc	
APPR	apprehensive	appr	v-tam	
ATT	attenuative	att	deriv-v	
AUD	auditive	aud	v-tam	
AUG	augmentative	aug	deriv-n	
CAP	captative	cap	deriv-v	
CAR	caritive	car	deriv-misc	
CAUS	causative	caus	deriv-v	
CNG	connegative	cng	v-nfin	
CO	coaffix	co	misc	
COND1	conditional converb 1	cond,cond1	v-nfin,v-nfin	
COND2	conditional converb 2	cond,cond2	v-nfin,v-nfin	
CTR	contrastive	ctr	disc	lexical
CVB	converb	cnvb	v-nfin	
CVB.IMM	immediate converb	cnvb,imm	v-nfin,v-nfin	
DIM1	diminutive 1	dim,dim1	deriv-n,deriv-n	
DIM2	diminutive 2	dim,dim2	deriv-n,deriv-n	
DIM3	diminutive 3	dim,dim3	deriv-n,deriv-n	
DIM4	diminutive 4	dim,dim4	deriv-n,deriv-n	
DIM5	diminutive 5	dim,dim5	deriv-n,deriv-n	
DIM6	diminutive 6	dim,dim6	deriv-n,deriv-n	
DIM7	diminutive 7	dim,dim7	deriv-n,deriv-n	
DIM8	diminutive 8	dim,dim8	deriv-n,deriv-n	
DRV	unspecified derivation	drv	deriv-v	
DST	destinative	dst	n-misc	
DST.PL	destinative, plural number	dst,pl	n-misc,n-num	
DSTR	distributive	dstr	deriv-misc	
DU	dual number	du	n-num	
DUB1	dubitative 1	dub1	v-tam	

Gloss	Description	Tsakorpus grammar tags	Tsakorpus category	Comment
DUB2	dubitative 2	dub2	v-tam	
DUR	durative	dur	v-tam	
DYA	dyadic (connective-reciprocal)	dya	deriv-n	
EMPH	emphatic marker	emph	disc	lexical
EP	epenthetic element	ep	misc	
ESS	essive case	ess	n-case	
EVID	evidential	evid	v-tam	
EVID.FUT	evidential, future tense	evid,fut	v-tam,v-tam	
EVID.PRS	evidential, present tense	evid,prs	v-tam,v-tam	
EVID.PST	evidential, past tense	evid,pst	v-tam,v-tam	
EXCL	exclamative	excl	disc	lexical
FRQ	frequentative	frq	v-tam	
FUT	future tense	fut	v-tam	
GEN	genitive case	gen	n-case	
HAB	habitual	hab	v-tam	
HORT	hortative	hort	v-tam	
IMP	imperative	imp	v-tam	
INCH	inchoative	inch	v-tam	
INDF	indefinite marker	indf	disc	lexical
INF	infinitive	inf	v-nfin	
INSTR	instrumental nominal	instr	deriv-n	
INT	intensifier	int	deriv-v	
INTER	interrogative	inter	v-tam	
INTERJ	unspecified interjection	interj	disc	lexical
IPFV	imperfectivizer	ipfv	v-tam	
JUSS	jussive	juss	v-tam	
LAT	lative case	lat	n-case	
LAT.ADV	lative case of adverbials	lat,advcase	n-case,n-misc	
LIM	limitative	lim	deriv-misc	
LOC	locative case	loc	n-case	
LOC.ADV	locative case of adverbials	loc,advcase	n-case,n-misc	
LOCN	location nominal	locn	deriv-n	
LOCZ	locativizer	locz	deriv-misc	
MD	middle conjugation	md	persnum-conj	only in combinations
MOD	modifier	mod	deriv-n	
MOM	momentaneous	mom	v-tam	
NAR	narrative	nar	v-tam	
NEC	necessitative	nec	v-tam	
NEG	negation	neg	neg	lexical
NEG.EMPH	emphatic negative	neg,emph	neg,neg	lexical
NEG.IMP	negative imperative	neg,imp	neg,neg	lexical
NMLZ	nominalization	nmlz	v-nfin	
NOM	nominative case	nom	n-case	only in combinations
NSG.O	objective conjugation, nonsingular object	objc,objnsg	persnum-conj, persnum-num	only in combinations
O	objective conjugation	objc	persnum-conj	only in combinations
OBL	oblique case	obl	n-case	only in combinations
OPT	optative	opt	v-tam	
ORD	ordinal numeral	ord	deriv-misc	
PASS	passive participle	pass	deriv-v	
PEJ	pejorative	pej	deriv-n	
PL	plural number	pl	n-num	
PLC	placeholder stem	plc	misc	
POSSV	possessive verbalizer	possv	deriv-misc	

Gloss	Description	Tsakorpus grammar tags	Tsakorpus category	Comment
PRO1DU	personal pronoun, 1 person dual	pers,pn1,pndu	pers,persnum-person, persnum-number	lexical
PRO1PL	personal pronoun, 1 person plural	pers,pn1,pnpl	pers,persnum-person, persnum-number	lexical
PRO1SG	personal pronoun, 1 person singular	pers,pn1,pnsg	pers,persnum-person, persnum-number	lexical
PRO2DU	personal pronoun, 2 person dual	pers,pn2,pndu	pers,persnum-person, persnum-number	lexical
PRO2PL	personal pronoun, 2 person plural	pers,pn2,pnpl	pers,persnum-person, persnum-number	lexical
PRO2SG	personal pronoun, 2 person singular	pers,pn2,pnsg	pers,persnum-person, persnum-number	lexical
PRO3DU	personal pronoun, 3 person dual	pers,pn3,pndu	pers,persnum-person, persnum-number	lexical
PRO3PL	personal pronoun, 3 person plural	pers,pn3,pnpl	pers,persnum-person, persnum-number	lexical
PRO3SG	personal pronoun, 3 person singular	pers,pn3,pnsg	pers,persnum-person, persnum-number	lexical
PROB	probabilitive	prob	v-tam	
PROG	progressive	prog	v-tam	
PROL	prolative case	prol	n-case	
PROP	proprietive	prop	deriv-misc	
PROS	prospective	pros	v-tam	
PRS	present tense	prs	v-tam	only in combinations
PST	past tense	pst	v-tam	
PTCP	participle	ptcp	v-nfin	only in combinations
PTCP.ABES	abessive participle	ptcp,abes	v-nfin,v-nfin	
PTCP.FUT	future participle	ptcp,fut	v-nfin,v-tam	
PTCP.NEC	necessitative participle	ptcp,nec	v-nfin,v-tam	
PTCP.PRS	present participle	ptcp,prs	v-nfin,v-tam	
PTCP.PST	past participle	ptcp,pst	v-nfin,v-tam	
PTCP.RES	resultative participle	ptcp.res	v-nfin	
RAR	raritive	rar	deriv-misc	
REP	reportative	rep	deriv-v	
S	subjective conjugation	subjc	persnum-conj	only in combinations
SEL	selective	sel	deriv-misc	
SG	singular number	sg	n-num	only in combinations
SG.O	objective conjugation, singular object	objc,objsg	persnum-conj, persnum-num	only in combinations
SIM	similative	sim	deriv-misc	
SPEC	speculative	spec	v-tam	
SPEC.FUT	speculative, future tense	spec,fut	v-tam,v-tam	
SPEC.PRS	speculative, present tense	spec,prs	v-tam,v-tam	
SPEC.PST	speculative, past tense	spec,pst	v-tam,v-tam	
SUP	supine	sup	v-nfin	
TOP	topical marker	top	disc	lexical
TR	transitivizer	tr	deriv-v	
TRL	translative	trl	deriv-misc	
VBLZ	verbalizer	vblz	deriv-misc	
VOC	vocative case	voc	n-case	
%	uncertain categories	uncert	misc	as part of a gloss, e.g. “%DRV”
%REP	possibly reportative	uncert,rep	misc,deriv-v	
%SPEC	possibly speculative	uncert,spec	misc,v-tam	
%%	unknown morph	unkn	misc	

Table 34. Alphabetical list of Tsakorpus grammar tags

Tsakorpus grammar tag	Tsakorpus category	Gloss	Description	Comment
abl	n-case	ABL, ABL.ADV	ablative case	
acc	n-case	ACC	accusative case	
act	deriv-n	ACT1, ACT2, ACT3, ACT4	action nominal (any)	
act1	deriv-n	ACT1	action nominal 1	
act2	deriv-n	ACT2	action nominal 2	
act3	deriv-n	ACT3	action nominal 3	
act4	deriv-n	ACT4	action nominal 4	
adjz	deriv-misc	ADJZ	adjectivizer	
advcase	n-misc	ABL.ADV, LAT.ADV, LOC.ADV	case of adverbials	
advz	deriv-misc	ADVZ	adjverbializer	
appr	v-tam	APPR	apprehensive	
att	deriv-v	ATT	attenuative	
aud	v-tam	AUD	auditive	
aug	deriv-n	AUG	augmentative	
cap	deriv-v	CAP	captative	
car	deriv-misc	CAR	caritive	
caus	deriv-v	CAUS	causative	
cng	v-nfin	CNG	connegative	
cnvb	v-nfin	CVB, CVB.IMM	converb	
co	misc	CO	coaffix	
cond	v-nfin	COND1, COND2	conditional converb	
cond1	v-nfin	COND1	conditional converb 1	
cond2	v-nfin	COND2	conditional converb 2	
ctr	disc	CTR	contrastive	lexical
dim	deriv-n	DIM1, DIM2,… DIM8	diminutive (any)	
dim1	deriv-n	DIM1	diminutive 1	
dim2	deriv-n	DIM2	diminutive 2	
dim3	deriv-n	DIM3	diminutive 3	
dim4	deriv-n	DIM4	diminutive 4	
dim5	deriv-n	DIM5	diminutive 5	
dim6	deriv-n	DIM6	diminutive 6	
dim7	deriv-n	DIM7	diminutive 7	
dim8	deriv-n	DIM8	diminutive 8	
drv	deriv-v	DRV	unspecified derivation	
dst	n-misc	DST, DST.PL	destinative	
dstr	deriv-misc	DSTR	distributive	
du	n-num	DU	dual number	
dub	v-tam	DUB1, DUB2	dubitative	
dub1	v-tam	DUB1	dubitative 1	
dub2	v-tam	DUB2	dubitative 2	
dur	v-tam	DUR	durative	
dya	deriv-n	DYA	dyadic (connective-reciprocal)	
emph	disc	EMPH, NEG.EMPH	emphatic marker	lexical
ep	misc	EP	epenthetic element	
ess	n-case	ESS	essive case	
evid	v-tam	EVID, EVID.FUT, EVID.PRS, EVID.PST	evidential	
excl	disc	EXCL	exclamative	lexical
frq	v-tam	FRQ	frequentative	
fut	v-tam	FUT, PTCP.FUT, SPEC.FUT	future tense	
gen	n-case	GEN	genitive case	
hab	v-tam	HAB	habitual	

Tsakorpus grammar tag	Tsakorpus category	Gloss	Description	Comment
hort	v-tam	HORT	hortative	
imm	v-nfin	CVB.IMM	immediate converb	
imp	v-tam	IMP, IMP.1SG, IMP.1DU, IMP.1PL,… IMP.3PL, IMP.2SG.SG.O, IMP.2DU.NSG.O,… NEG.IMP	imperative	
inch	v-tam	INCH	inchoative	
indf	disc	INDF	indefinite marker	lexical
inf	v-nfin	INF	infinitive	
instr	deriv-n	INSTR	instrumental nominal	
int	deriv-v	INT	intensifier	
inter	v-tam	INTER	interrogative	
interj	disc	INTERJ	unspecified interjection	lexical
ipfv	v-tam	IPFV	imperfectivizer	
juss	v-tam	JUSS	jussive	
lat	n-case	LAT, LAT.ADV	lative case	
lim	deriv-misc	LIM	limitative	
loc	n-case	LOC, LOC.ADV	locative case	
locn	deriv-n	LOCN	location nominal	
locz	deriv-misc	LOCZ	locativizer	
md	persnum-conj	MD	middle conjugation	only in combinations
mod	deriv-n	MOD	modifier	
mom	v-tam	MOM	momentaneous	
nar	v-tam	NAR	narrative	
nec	v-tam	NEC, PTCP.NEC	necessitative	
neg	neg	NEG, NEG.EMPH, NEG.IMP	negation	lexical
nmlz	v-nfin	NMLZ	nominalization	
nom	n-case	NOM.SG, NOM.DU, NOM.PL, NOM.3SG, NOM.3DU, NOM.3PL,…	nominative case	only in combinations
objc	persnum-conj	3SG.O, 1PL.SG.O, 1PL.NSG.O,…	objective conjugation	only in combinations
objnsg	persnum-num	1SG.NSG.O, 1PL.NSG.O,…	nonsingular object	only in combinations
objsg	persnum-num	1SG.SG.O, 1PL.SG.O,…	singular object	only in combinations
obl	n-case	OBL.1SG, OBL.1DU, OBL.1PL,… OBL.3PL	oblique case	only in combinations
opt	v-tam	OPT	optative	
ord	deriv-misc	ORD	ordinal numeral	
pass	deriv-v	PASS	passive participle	
pej	deriv-n	PEJ	pejorative	
pers	pers	PRO1SG, PRO1DU, PRO1PL,… PRO3PL	personal pronoun	
pl	n-num	PL, DST.PL	plural number	
plc	misc	PLC	placeholder stem	
pn1	persnum-person	1SG, 1DU, 1PL	1 person	
pn2	persnum-person	2SG, 2DU, 2PL	2 person	
pn3	persnum-person	3SG, 3DU, 3PL	3 person	
pndu	persnum-number	1DU, 2DU, 3DU	dual number	
pnpl	persnum-number	1PL, 2PL, 3PL	plural number	
pnsg	persnum-number	1SG, 2SG, 3SG	singular number	
poss	poss	NOM.SG.1SG, …	possessive	
possv	deriv-misc	POSSV	possessive verbalizer	
prob	v-tam	PROB	probabilitive	
prog	v-tam	PROG	progressive	

Tsakorpus grammar tag	Tsakorpus category	Gloss	Description	Comment
prol	n-case	PROL	prolative case	
prop	deriv-misc	PROP	proprietive	
pros	v-tam	PROS	prospective	
prs	v-tam	EVID.PRS, PTCP.PRS, SPEC.PRS	present tense	only in combinations
pst	v-tam	PST, EVID.PST, PTCP.PST, SPEC.PST	past tense	
ptcp	v-nfin	PTCP.ABES, PTCP.FUT, PTCP.NEC, PTCP.PRS, PTCP.PST, PTCP.RES	participle	only in combinations
abes	v-nfin	PTCP.ABES	abessive participle	
res	v-nfin	PTCP.RES	resultative participle	
rar	deriv-misc	RAR	raritive	
rep	deriv-v	REP	reportative	
sel	deriv-misc	SEL	selective	
sg	n-num	ABL.SG, NOM.SG, PROL.SG,…	singular number	only in combinations
sim	deriv-misc	SIM	similative	
spec	v-tam	SPEC, SPEC.FUT, SPEC.PRS, SPEC.PST	speculative	
subjc	persnum-conj	1SG.S, 2SG.S,… 3PL.S,…	subjective conjugation	only in combinations
sup	v-nfin	SUP	supine	
top	disc	TOP	topical marker	lexical
tr	deriv-v	TR	transitivizer	
trl	deriv-misc	TRL	translative	
uncert	misc	%	uncertain categories	as part of a gloss, e.g. “%DRV”
unkn	misc	%%	unknown morph	
vblz	deriv-misc	VBLZ	verbalizer	
voc	n-case	VOC	vocative case	

Appendix A5. Part-of-speech and morphological category tags (tier mc, ps)

Table 35. Part-of-speech tags (tiers mc, ps)

Tag	Description
adj	adjective
adv	adverb
aux	auxiliary
conj	conjunction
cop	copula
cvb	converb
dem	demonstrative pronoun
intj	interjection
interrog	interrogative pro-form
n	noun
npr	proper noun
num	numeral
pp	postposition
pro	pronoun
ptcl	particle
ptcp	participle
quant	quantifier
reln	relational noun
v	verb

Table 36. Morphological categories tags (tier mc)

Label	Description
adv:(case)	Adverbial case marker
n:(case)	Nominal case marker
n:(case.poss)	Nominal case and possessive marker
n:(dst)	Nominal destinative marker
n:(ins)	Nominal co-affix
n:(num)	Nominal number marker
n:(num.poss)	Nominal number and possessive marker
n:(poss)	Nominal possessive marker
pp:(case)	Postpositional case marker
v:(conj)	Verbal marker of non-singular object
v:(evid)	Verbal evidential marker
v:(ins)	Verbal co-affix
v:(mood)	Verbal mood marker
v:(mood:pn)	Verbal mood and person-number marker
v:(tense)	Verbal tense marker
v:infl	Verbal marker of uninflected finite forms
v:nfin	Verbal marker of uninflected non-finite forms
v:pn	Verbal person-number marker