



# INEL Selkup corpus

## User documentation

Svetlana Orlova, Maria Brykina, Alexandre Arkhipov, 25.06.2020

### 1. Introduction

#### 1.1. Objective of the corpus

The present corpus of the Selkup language has been developed as part of the long-term research project INEL (“Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages”). It brings to a wide linguistic audience the contents of the archive of the Russian linguist Angelina Kuzmina, who worked extensively on different Selkup dialects in 1960s and 1970s. Most part of her archive remained unpublished until present, although some texts were published by Kuzmina herself and some more by other researchers (see References). The corpus makes possible typologically aware corpus-based grammatical research on the Selkup language and expands the documentation of the lesser described indigenous languages of Northern Eurasia.

#### 1.2. Selkup language

##### 1.2.1. Description

Selkup belongs to the Samoyedic branch of the Uralic language family. It is spoken in the Western Siberia, between two rivers — the Ob and the Yenisei — in Yamalo-Nenets AO, Krasnoyarsk Krai and Tomsk Oblast. Despite a vast geographical extent, the Selkup population probably never reached high numbers due to the features of the natural environment, making the whole area traditionally inhabited by Selkups extremely difficult to travel: the main transport ways are waterways, and most of the inhabited places are fairly isolated from one another.

Selkup is at present critically endangered, and though according to the census 2010 [VPN 2010] there are 3,649 people identifying themselves as Selkups, the language is spoken or understood only by a few dozen people. Most of them are native speakers of Northern dialects (see 1.2.3), while the other varieties of Selkup are actually almost extinct.

Selkup 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 Selkup nouns are number, case, possession, while verbs carry markers of several aspectual categories, mood and evidentiality, tense, as well as person and number agreement.

##### 1.2.2. Language Codes

ISO-639-3 code: **sel**

Glottolog code: **selk1253**

##### 1.2.3. Dialectal subdivisions

Classification of Selkup dialects is an area of debate. Two, three or four larger dialect groups are distinguished by different researchers. Most notably, Northern Selkup dialects are clearly distinct from all the others, which are sometimes grouped together, but often subdivided into Central and Southern dialects. Furthermore, Ket dialects, which otherwise are included in the Southern group, can be treated as a group on their own. Thus, in Russian historical, ethnographic and formerly also in linguistic tradition only Northern and Southern Selkup are differentiated

(see [Glushkov et al. 2013]). On the other hand, Helimski [1998] considers the Ket<sup>1</sup> dialects as a fourth dialect group. The names for the particular dialects mostly refer to names of the rivers along which the speakers used to dwell. River Taz in the north flows into the Kara Sea through the Gulf of Ob, and Turukhan is a tributary of Yenisey, while the rivers in the central and southern area are part of the Ob river system.

The investigations of Selkup dialects are confronted with difficulties because of (i) lack of systematic data on many varieties, most of which are already extinct or on the verge of extinction, (ii) well-known but poorly documented sporadic migrations of Selkups between different Selkup-speaking settlements during their lives, obscuring correlations between geographic areas or specific settlements and linguistic features.

In this release of the corpus, the threefold distinction is used in metadata with further subdivisions (boldface marks those dialects for which texts are present in the corpus):

Dialect group	Dialects	Subdialects	Comments
<b>Northern</b>			= Taz-Turukhan
	<b>Taz</b>	<b>Middle Taz, Upper Taz</b>	
	<b>Upper Tolka</b>		= Laryak; settlement Tolka in Purovsky district on the river Tolka, tributary of Taz
	<b>Baikha</b>		= Turukhan
	Yelogui		
<b>Central</b>			
	<b>Narym</b>		one speaker (KFN) is here classified as mixed Narym/Tym dialect
	<b>Tym</b>		
	Vasyugan		
	Vakh		
<b>Southern</b>			
	<b>Ob</b>	<b>Upper Ob, Middle Ob</b>	
	<b>Chaya</b>		
	Chulym		
	<b>Ket</b>	<b>Lower Ket, Middle Ket and Upper Ket</b>	

Different groups of Selkups had extensive contacts with different populations and their languages, including Khanty, Ket, Evenki, Nenets and Turkic (Chulym Turkic and Siberian Tatar) [Helimski 1998].

### 1.3. Archiving

The corpus comprises source media files (whenever available), annotated transcripts in EXMARALDA<sup>2</sup> transcript formats and metadata descriptions in EXMARALDA Coma format (see 2.7 and 2.9 for details).

The publication and archiving is performed in collaboration with the Hamburg Centre for Language Corpora (HZSK).<sup>3</sup>

The corpus is freely available under open-access conditions with Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0).<sup>4</sup>

### 1.4. Citation

The corpus is to be cited as follows:

Brykina, Maria; Orlova, Svetlana; Wagner-Nagy, Beáta. 2020. "INEL Selkup Corpus." Version 1.0. Publication date 2020-06-30. <http://hdl.handle.net/11022/0000-0007-E1D5-A>. Archived in Hamburger Zentrum für Sprachkorpora.

<sup>1</sup> The name of Ket dialects of Selkup is due to the river of Ket` (Кеть), not to be confounded with the Ket language, which comes from the word meaning 'man'.

<sup>2</sup> <http://exmaralda.org/en/>, last access: 18.12.2018.

<sup>3</sup> <https://corpora.uni-hamburg.de/hzsk/en>, last access: 18.12.2018.

<sup>4</sup> <https://creativecommons.org/licenses/by-nc-sa/4.0/>, last access: 18.12.2018

In: Wagner-Nagy, Beáta; Arkhipov, Alexandre; Ferger, Anne; Jettka, Daniel; Lehmberg, Timm (eds.). The INEL corpora of indigenous Northern Eurasian languages.

## 1.5. Project members

### Project summary information

The INEL Selkup 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 research was carried out at the Institute for Finno-Ugric/Uralic Studies (IFUU) of the Hamburg University (UHH).

The technical infrastructure is based on groundwork of the Hamburg Centre for Language Corpora (HZSK).

The project homepage can be visited at: <https://inel.corpora.uni-hamburg.de/>.

### Project leader

Prof. Dr. Beáta Wagner-Nagy (IFUU, Hamburg University)

### Researchers

Dr. Alexandre Arkhipov, Research coordinator

Dr. Maria Brykina (February – July 2016, August 2017 – December 2018)

Dr. Svetlana Orlova (August 2016 – August 2018)

Prof. Dr. Beáta Wagner-Nagy

Josefina Budzisch, M.A.

Dr. Chris Lasse Däbritz

Hannah Wegener, M.A.

Contributions of particular researchers are acknowledged in more detail in the metadata to the corpus (see 2.1.3).

### Developers

Timm Lehmberg, M.A., Technical coordinator

Daniel Jettka, M.A.

Anne Ferger, M.A. (since April 2017)

Niko Partanen, M.A. (February 2016 – March 2017)

### Student assistants

Hannes Klitzing (September – December 2016)

Olesya Degtyareva (October 2016 – December 2017)

Felix Templin (April 2016 – June 2018)

Gerrit Jawinsky (April 2017 – June 2018)

Ozan Özdemir (August 2018 – December 2018)

Theodor Hey (April – September 2019)

Jacqueline Krieg (since November 2019)

## 1.6. Acknowledgements

### Funding

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

## Data analysis

Sound materials of Angelina Kuzmina were transcribed and translated by native speakers of Selkup:

Irina Anatolyevna Korobejnikova, written transcription and Russian translation of audio in Central and Southern dialects

Natalya Platonovna Izhenbina, written transcription and Russian translation of audio in Southern dialects

Svetlana Nikitichna Sankevich (Kunina), oral transcription and Russian translation of audio in Northern dialects

Evgeniya Sergeevna Smorgunova (Irikova), oral and written transcription and Russian translation of audio in Northern dialects

Valentina Vladimirovna Tamelkina, oral transcription and Russian translation of audio in Northern dialects

Part of translations into English were done by Anna Fenyvesi.

Part of translations into German were done by Aleksandra Vladimirovna Bajdak, Elena Aleksandrovna Kryukova and Ralph Reindler.

### 1.7. New in release 1.0

- The corpus now contains 264 texts from 74 speakers, representing the dialects of Middle Taz, Upper Tolka, Baikha (Northern), Narym and Tym (Central), Upper and Middle Ob, Chaya, Upper and Middle Ket (Southern). These contain 7887 sentences and 42466 words in total.
- Many texts have been provided with annotations for syntactic functions and semantic roles.
- Corrections to audio transcriptions, glossing and other annotations.

## 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 Selkup, with few instances of code-switching into Russian (slightly more of them in the audio materials).

The same content can however be represented in more than one form of transcription, depending on the source. There is always only one main transcription tier (per speaker), using the common INEL transcription style (see 2.10.2).

#### 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**).

For texts from the written archive, original translation into Russian (usually a literal word-by-word translation, often incomplete) is given as provided in the manuscripts (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 language of metadata is English; Russian spellings of the personal names and place names are also provided in communications and speaker metadata. On transliteration of names, see 2.9.1.

## 2.2. Sources

The INEL Selkup corpus originates from the archive of A.I. Kuzmina representing the materials of her field work with native speakers of different Selkup varieties in 1960s and 1970s. A detailed description of the archive by E.A. Helimsky and N.A. Tuchkova is published in [Tučkova, Helimski 2010] (in Russian and in German).

The Kuzmina archive preserved at IFUU (Hamburg) includes handwritten notebooks bound in volumes (30 volumes, 357 notebooks) and a collection of sound recordings digitized from her reel-to-reel tapes (ca. 50 tapes). The corpus includes both written and audio data.

The written part of the Selkup archive of Angelina Kuzmina contains a large amount of texts (both original Selkup texts and some translations from Russian) and translations of individual sentences (from Russian into Selkup). The texts are transcribed in Selkup with interlinear (word-by-word) translation into Russian. Under or near the utterances some comments on linguistic or ethnolinguistic details can be found. Apart from these, the archive also contains translated sentences, lexical data, grammatical paradigms and other kinds of data, which were not included in the corpus.

Angelina Kuzmina also noted some general information about native speakers, time and place of collecting the data (her notes usually precede the linguistic material in notebooks). These metadata are included in the corpus metadata.

## 2.3. Content

The corpus contains texts/transcripts of various genres, which are broadly classified as folklore, narrative, conversation and song; while not being a separate genre, translations are classified apart from the other genres, for their language differs in some respects from the original Selkup texts.

## 2.4. Selection

From the entire body of Kuzmina archive, only those data have been selected for the corpus which represent coherent texts (although one or two are more likely collections of loosely connected sentences).

## 2.5. Corpus size

This release of the corpus contains 264 transcripts (57 folklore, 183 narrative, 11 translations, 11 songs, 2 conversations) of 74 speakers, with a total of 7 887 sentences and 42 466 words. 63 texts come from 35 speakers of the Northern dialect group (Middle Taz, Upper Tolka, Baikha), 33 texts come from 10 speakers of the Central dialect group (Tym, Narym), 168 texts come from 30 speakers of the Southern dialect group (Upper and Middle Ob, Chaya and Ket). Among them, 100 texts come from a single speaker of Chaya dialect (PVD).

## 2.6. Naming Conventions

### 2.6.1. Name of the corpus

The name of the corpus is INEL Selkup Corpus.

### 2.6.2. Orthography conventions in the corpus

For transliteration of Cyrillic names in the metadata, see 2.9.1.

### INEL transcription system

The transcription of Selkup adopted in the INEL project is based on FUT (Finno-Ugric Transcription system). The project transcription is represented in tiers **ts** (Text (Sentence)) and **tx** (Text (Word)). These tiers represent the text in a Latin-based transcription which was developed in the project. Vowel length is marked as <V:>, i.e. the “Modifier Letter Triangular Colon” character after the vowel. Palatalization is marked as <C’>, i.e. the consonant symbol followed by the “Modifier Letter Apostrophe” character. In the corpus the Charis SIL font is used.

The INEL transcription for Northern Selkup dialects is largely a phonemic one. For the set of phonemes used in the transcription see [Appendix 1](#). The INEL transcription for Southern and Central Selkup dialects is not strictly phonemic and follows the transcription of A.I. Kuzmina, omitting most of her diacritics.

## Original transcriptions and their conversion into INEL transcription

While transcribing the Selkup data, Angelina Kuzmina used in her notebooks a transcription system developed by A. P. Dulson based on the Cyrillic alphabet. Besides the 32 Cyrillic letters (except *ë*) it also makes use of some Latin characters (**k, j, l, w, y, ə, ɛ**), additional diacritics ( *¨ ° ~ ´ ´ ˘ ˘ ˘ ˘ ˘ ˘* ) and stress markers ( *˘ ˘* ). One character may bear more than one diacritic symbol. The original Kuzmina's transcriptions are provided in the **st** tier (Source transcription).

This original transcription was first automatically converted into Latin alphabet (see tier **stl**, Source transcription (Latin)). However, in order to reduce variation and render the interlinearization process easier, this transcription was further adjusted. For the Northern dialect some amendments were made according to the phonemic transcription principle (e.g. the voiced consonants were replaced with their voiceless counterparts), or else with consideration of the particularities of Russian perception of Selkup pronunciation (since the author of manuscripts was a native speaker of Russian). This amended Latin transcription is the main transcription used in tiers **ts** (Text (Sentence)) and **tx** (Text (Word)).

The changes to the transcription made automatically and manually are summarized in [Appendix 2](#).

## Capitalization and punctuation

In original manuscripts of Kuzmina there were often no punctuation marks and capital letters at all, so they have been added while glossing the texts in FLEx.

### 2.6.3. Folder structure

The entire corpus is contained in the folder "SelkupCorpus" which has the following files and subfolders.

Folders with text transcripts, organized by genre:

- "flk" (folklore texts)
- "nar" (narrative texts)
- "song" (texts of songs)
- "conv" (conversations)
- "transl" (texts translated into Selkup from Russian)

Each of these genre folders contains one further subfolder per each communication, named identically to the communication name (see 2.6.6.1). Each communication folder contains several files with the same filename identical to the communication name, and different extensions according to the file type (see 2.7 for details on file formats):

- annotated transcript in EXMARaLDA EXB and EXS formats (\*.exb, \*.exs)
- scanned manuscript pages from the Kuzmina archive, as well as of published versions of the text (if available), in PDF (\*.pdf) (for texts with written source)
- sound file with the digitized recording from the Kuzmina archive, in WAV (\*.wav) (for texts with audio source)

Supplementary folder:

- "documentation" (contains user documentation)
- "corpus-utilities" (contains annotation panel for use in EXMARaLDA Partitur Editor and segmentation settings)

Individual files:

- "selkup.coma" (main metadata file)
- "coma\_overview.html" (a browser-readable overview of the main metadata file)

### 2.6.4. Transcripts

The names of the transcript files have the structure Speaker\_DateOfRecording\_Title\_Genre, i.e. they have the same name as the respective communication in the metadata (see 2.6.6.1 for details). The file name extensions are .exb and .exs for the basic and segmented transcript files respectively (see 2.7.1).

## 2.6.5. Media

The names of the sound files have the structure Speaker\_DateOfRecording\_Title\_Genre, i.e. the same name as the respective communication in the metadata (see 2.6.6.1 for details). The file name extension is .wav.

## 2.6.6. Metadata

The main metadata file for the corpus is the selkup.coma file stored in the main corpus folder (EXMARaLDA Coma format; see 2.7.2 for details). It contains the metadata on speakers and on individual communications (texts).

### 2.6.6.1. Names of communications

The codes of the communications which are used as their IDs throughout the corpus are composed of the following components: speaker code (see 2.6.6.2); year of recording; communication short title, genre abbreviation. These components are joined by underscore (“\_”).

The exact date is mentioned in the communication code if known, in the format YYYYMMDD. If the day or both the day and the month are unknown, they are omitted (thus YYYYMM or YYYY). If the year of recording is only approximate or altogether unknown, a placeholder character "X" is used to fill the missing digits (e.g., “196X”). In the communication metadata, only the year of recording is specified.

The communication short title is a (possibly shortened) version of the English title, spelled without spaces, dashes or other non-letter characters, with all initial capitals. This English title is usually a translation of the Russian title, which is generally given by the corpus creators, however in some cases the titles are provided in the manuscript or follow existing publications.

The genre abbreviation can have one of the values “flk” (folklore), “nar” (narrative), “song” (song), “transl” (translation) and “conv” (conversation).

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

**Communication code:** TVP\_1965\_ThreeBrothersLapta\_flk

**Speaker code:** TVP (Ton`kin, Viktor Petrovich)

**Date of recording:** 1965

**Short title:** ThreeBrothersLapta (i.e. “Three brothers Lapta”)

**Genre:** flk (folklore)

A number of texts are found both in the audio recordings and in Kuzmina’s notebooks. In most cases the written transcript by Kuzmina is far from being identical to the audio, rather representing another, more or less close, version of the same story. Such pairs of texts are given the same title, ending with “1” for the written version and ending with “2” for the audio version retranscribed within the current project. The year code can still be different if it is not clear from the available metadata if it was originally recorded and transcribed in the same year, cf. KFN\_1965\_DrunkBear1\_nar and KFN\_196X\_DrunkBear2\_nar.

### 2.6.6.2. 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 (simplified). Most commonly, a code is thus composed of three initial capital letters, e.g. “TVP” stands for Ton`kin, Viktor Petrovich (Тонькин, Виктор Петрович). If the patronymic is not noted by Kuzmina, only initials of the family name and of the first name are used, e.g. “KR” for Kunina Rita (Кунина, Рита). If a code is already assigned to a different speaker, including from another language in the INEL project, additional letters are used from one or more of the name parts, e.g. KNM for Kunin, Nikita Mixajlovich (Кунин, Никита Михайлович) and KNiM for Kunin Nikolaj Mixajlovich (Кунин, Николай Михайлович).

## 2.6.7. Abbreviations

Data collectors and editors

KuAI: Kuz`mina, Angelina Ivanovna

KNS: Karsavin, Nikolaj (self-transcription of one text)

## Project members

AAV: Arkhipov, Alexandre

BJ: Budzisch, Josefina

BrM: Brykina, Maria

DCh: Däbritz, Chris Lasse

OSV: Orlova, Svetlana

WH: Wegener, Hannah Christine

WNB: Wagner-Nagy, Beáta

## Student assistants

DO: Degtyareva, Olesya

KH: Klitzing, Hannes

KJ: Krieg, Jacqueline

## Language consultants (transcription and translation)

IES: Smorgunova (Irikova), Evgeniya Sergeevna

INP: Izhenbina, Natalya Platonovna

KoIA: Korobejnikova, Irina Anatolyevna

KSN: Sankevich (Kunina), Svetlana Nikitichna

TVV: Tamel'kina, Valentina Vladimirovna

## Other contributors

BAV: Bajdak, Aleksandra Vladimirovna

FA: Fenyvesi, Anna

HA: Harder, Anja

JF: Jark, Florian

KrE: Kryukova, Elena Aleksandrovna

ReR: Reindler, Ralph

TN: Tuchkova, Natalya Anatolyevna

## 2.7. Technical formats

### 2.7.1. Transcripts

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

### 2.7.2. Metadata

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

### 2.7.3. Media

For texts with audio source, sound files are provided in Linear PCM WAVE format (file extension “.wav”) mono, with 44 100 Hz sampling frequency and 16 bit depth. However it should be noted that it is not their native format, since Kuzmina’s recordings were originally analog and later digitized and stored on audio CDs (see 2.8.2).

For texts with written source, corresponding pages scanned from Kuzmina’s manuscripts are provided in PDF format (file extension “.pdf”).

#### 2.7.4. Other data

No other data types are provided with the corpus.

### 2.8. Workflow of the source files

#### 2.8.1. Transcripts

The workflow was different depending on the source of the text.

Texts from the manuscripts of Kuzmina’s archive were typed manually into text files, containing the original Cyrillic transcription by Angelina Kuzmina (**st**), as well as her original (word-by-word) Russian translation (**ltr**) and occasional original notes (**nto**). Afterwards, the Cyrillic transcription was automatically latinized (**stl**) and manually adapted to produce the INEL transcription (**ts**) – see 2.6.2 above. The resulting text was then imported into *SIL Fieldworks Language Explorer* (FLEX)<sup>5</sup> for glossing.

Texts from the digitized Kuzmina magnetic tapes were transcribed in fieldwork sessions with native speakers. The transcription was done in ELAN multimedia annotator<sup>6</sup>. The transcript files containing transcription (**ts**), literal Russian translation provided by native speakers (**ltr**), our free translation (**fr**) and occasional notes (**nt**), were saved in flextext format and imported into FLEX for glossing.

For all transcripts, the morphological analysis (interlinear glossing) was done in FLEX. This is when all the morpheme-level tiers were created (**mb**, **mp**, **ge**, **gr**, **mc**), as well as the part-of-speech tier (**ps**). For most texts, the **BOR** tier was also filled directly from the FLEX lexicon. Together with the glossing, Russian translation (**fr**) and notes (**nt**) were provided. Separate FLEX databases ("projects") were maintained for different dialects: Northern, Central, Southern except Ket, and Ket.

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

After that, all further annotating (and editing) is done in the EXMARaLDA Partitur-Editor (see also 2.10).

#### 2.8.2. Media

The original sound recordings were analog recordings made by Angelina Kuzmina in 1960s and 1970s on reel-to-reel tapes. They were digitized by G. Soldatova (Institute of Philology, SO RAN, Novosibirsk) in 2001 onto CDs, digital copies of which were used in the project. The audio quality was poor for many of the files which were therefore processed to improve intelligibility. The volume normalization and basic editing was performed in Sound Forge 12, and noise reduction / reverb reduction in Spectralayers 3.0. Even processed sound is sometimes of very poor quality, which is naturally reflected in the tentative character of transcriptions.

Published texts from Kuzmina’s archive were scanned and attached in PDF format to the metadata records of corresponding communications in the corpus manager (see 2.9.3).

#### 2.8.3. Metadata

The communication and speaker metadata were first extracted from the description of Kuzmina’s archive published in [Tučkova, Helimski 2010] and transferred into EXMARaLDA Coma, being also completed and cross-checked with the notes in Kuzmina’s manuscripts.

### 2.9. Metadata for the corpus

The metadata of the corpus are stored in *EXMARaLDA Coma* format. It is an XML-based format with separate interlinked descriptions for communications (texts; also analogous to IMDI “sessions”) and speakers. The main fields

---

<sup>5</sup> <https://software.sil.org/fieldworks/>, last access: 26.10.2017.

<sup>6</sup> <https://tla.mpi.nl/tools/tla-tools/elan/>, last access: 28.11.2017.

contained in the descriptions are listed in the following sections. This includes for example the location and date of a communication, but also information on which part of the processing and analysis was done by whom. Metadata about speakers contains mainly biographical data, but also basic data on language proficiency.

### 2.9.1. Naming conventions and content of the metadata

The general metadata about the whole corpus include the corpus name (“INEL Selkup Corpus”) and some basic metadata fields complying with the standards of DC (Dublin Core), OLAC (Open Language Archive Community) and HZSK (Hamburger Zentrum für Sprachkorpora).

The transliteration (romanized spelling) of Russian place names, except major cities, and personal names of native speakers follows the GOST 7.79–2000 System B transliteration standard. Elsewhere, common English spelling is used. Thus, the village name *Толька* is rendered as *Tol'ka*, but the dialect name as *Upper Tolka*; the settlement name *Туруханск* is spelled as *Turuxansk*, but the dialect of *Бауха* is referred to as *Baikha*.

### 2.9.2. Communication metadata

**Name:** The code which is assigned to the communication (see 2.6.6.1)

**Description:**

- **0a. Title:** Short title (in English)
- **0b. Title (RU):** Short title (in Russian)
- **1. Genre:** Abbreviation of the genre of the communication (flk = folklore, nar = narrative, song = song, transl = translation, conv=conversation).
- **2a. Recorded by:** Abbreviation of the person by whom the communication was recorded (see 2.6.8)
- **2b. Date of recording:** Here the date of recording is given (year only).
- **3a-c. Dialect group / Dialect / Subdialect:** If known, information on the dialect used by the speaker(s) is given here (see 1.2.3); if not, the field remains empty.
- **4. Speakers:** Code of the speaker
- **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
- **5d. Time-aligned by:** Code of the person who time-aligned the transcription (if done separately from the transcription)
- **6a-b. Processed by / Date of processing:** Who and when applied technical processing to the source files (e.g. noise reduction)
- **7a-c. Translation into Russian / English / German:** Code of the person who did the translation in question. For most of the written material, Angelina Kuzmina (KuAI) is mentioned as the first translator into Russian
- **8a. Glossed by:** Code of the person who did the morphological glossing
- **8b. Glosses checked:** Whether the glosses have been double-checked
- **9a-d. Annotation SeR / SyF / IST / BOR\_CS:** Codes of the persons who annotated the respective tiers (SeR; SyF; IST; BOR, BOR-Phon, BOR-Morph and CS; see 2.10.3)

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

- **Country:** All the communications originate from Russia
- **Region:** The current administrative region is indicated
- **Settlement:** The place of the recording

**Languages:**

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

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

- **1a. Archive (written):** For materials from the written part of the Kuzmina archive, their location in the archive is specified in the following format: **KA V[volume\_number]:B[book\_number]:[page\_range]**, e.g. **KA V02:B20:450-456** (Kuzmina archive, Volume 2, Book 20, pages 450-456). If a single text spans over multiple notebooks, the second and further notebooks are given after comma: **KA V02:B14:338-342, B15:343-355**. For texts from the audio archive which have a counterpart in the written archive with no or only partial correspondence in transcription, the reference to the written archive volumes is given in brackets (see field “Corresp. sound/written” below).
- **1b. Number of pages:** Number of manuscript pages in the archive
- **2. Corresp. sound/written:** If a text from the written archive has a counterpart in the sound recordings, the degree of correspondence in transcription is mentioned here (yes/no/partly).
- **3a. Published in:** If the text has been published, the publication reference is provided here
- **3b. 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.

### 2.9.3. Speaker metadata

Metadata about the speaker(s) taking part in a communication include on the one hand biographical information of the speaker and on the other hand information on his sociolinguistic background. The level of detail is determined by the information available from the manuscript archive. Many fields exist both in Russian (RU) and English (translated or transliterated) version. The following fields are defined:

**Signle:** Speaker code as defined in 2.6.6.2

**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. Vulgo (Sel. name):** Before getting Russian names, Selkups had their own names and principles of naming persons; if the Selkup name of a speaker is known, it is given here.
- **5a-b. Alternate names (EN, RU):** If alternate names (e.g. maiden name, short name/diminutive) or name spellings are found, they are given here

**Basic biographic data:** Here basic biographical data of the speaker are provided.

- **1a-b. Place of birth (EN, RU)**
- **2. Region**
- **3. Country:** Russia
- **4. Date of birth**
- **5. Date of death**
- **6a-b. Former residences (EN, RU):** If former residences prior to the work with the linguist are known, they are mentioned here
- **7a-b. Domicile (EN, RU):** Here the current (i.e. at the time of the recording) place of residence of the speaker is mentioned and, if known, date from which he or she started to live in this place
- **8a-b. Other information (EN, RU)**

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

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

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

- **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 / Names of grandparents**
- **6a-b. Family (EN, RU):** other family information

**Language documentation activities**

- **Informant of:** Here the linguist with whom the speaker worked is mentioned. It is normally Angelina Kuzmina (KuAI).

**Languages:** Here the language and dialect of the speaker is noted (*sel* – Selkup, *ket* – Ket); Russian (most commonly L2) is not mentioned.

**L1 (First language)**

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

**L2 (Second language)**

- **Language code:** Here the ISO code is given (*ket* – Ket).
- **Second language**

## 2.10. 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.

### 2.10.1. Tier layout

*Table 1. Tiers in EXMARaLDA files of INEL Selkup Corpus*

Tier label	Tier full name	Description	Unit	Optionality
ref	Reference	Text ID + sentence number	sentence	obligatory
st	Source transcription	Original phonetic transcription by A. I. Kuzmina (for texts from the written archive). Original transcription by native speakers (for some texts from the audio archive)	sentence	obligatory (written source); optional (audio source)
stl	Source transcription latinized	Original phonetic transcription by A. I. Kuzmina, converted into Latin alphabet and simplified (written archive). Original transcription by native speakers converted into Latin alphabet (for some texts from the audio archive). Original transcription by project members during work sessions with native speakers (for other texts from the audio archive).	sentence	obligatory
ts	Text (sentence)	Main transcription, adapted from stl tier for glossing and annotation.	sentence	obligatory

<b>Tier label</b>	<b>Tier full name</b>	<b>Description</b>	<b>Unit</b>	<b>Optionality</b>
<b>tx</b>	Text (word)	Main transcription segmented by word	word	obligatory
<b>mb</b>	Morpheme breaks	Morpheme breakdown of words (morphemes dash-separated for each word)	morph	obligatory
<b>mp</b>	Morphemes (lexical)	Lexical representation of morphemes (see notes in 2.10.3.3)	morph	obligatory
<b>ge</b>	Gloss (English)	Morpheme glosses (with lexical glosses in English)	morph	obligatory
<b>gr</b>	Gloss (Russian)	Morpheme glosses (with lexical glosses in Russian)	morph	obligatory
<b>mc</b>	Morphological category	Morphological category/part of speech for each morpheme	morph	obligatory
<b>ps</b>	Part of speech	Part of speech for each word	word	obligatory
<b>SeR</b>	Semantic role	Semantic (thematic) roles for major NPs	word / group of words	optional
<b>SyF</b>	Syntactic function	Syntactic functions for predicates and arguments	word / group of words	optional
<b>IST</b>	Information status	Information status for major NPs (given/new/accessible)	word	optional
<b>BOR</b>	Borrowing	Borrowings (source language and borrowing type)	word	optional
<b>BOR-Phon</b>	Borrowing phonology	Phonological adaptations in borrowings	word	optional
<b>BOR-Morph</b>	Borrowing morphology	Morphological adaptations in borrowings	word	optional
<b>CS</b>	Code switching	Code switching and calques (source language and type)	group of words	optional
<b>fr</b>	Free translation (Russian)	Free translation (Russian)	sentence	obligatory
<b>fe</b>	Free translation (English)	Free translation (English)	sentence	obligatory
<b>fg</b>	Free translation (German)	Free translation (German)	sentence	obligatory
<b>ltr</b>	Literal translation (Russian)	Original Russian translation, as provided in Kuzmina manuscripts (written archive). Original Russian translation, as provided by native speakers while transcribing (audio archive).	sentence	optional
<b>nt</b>	Notes	Notes from corpus developers	sentence	optional
<b>nto</b>	Notes (original)	Notes by Kuzmina / by native speakers while transcribing	sentence	optional

Here is an example of how a simple sentence looks like in the corpus:

Figure 1. A sample transcript showing the complete tier layout

<b>ref</b>	KuLP_1976_Hospital_nar.003 (001.003)					
<b>st</b>	мат пол'ницауыт 'иппысам 'нӧӧр и'ренты кунты.					
<b>stl</b>	mat pol'nicaqit ippisam no:kur irenti kunti.					
<b>ts</b>	Mat pol'nicaqit ippisam no:kur iränti kunti.					
<b>tx</b>	<b>Mat</b>	<b>pol'nicaqit</b>	<b>ippisam</b>	<b>no:kur</b>	<b>iränti</b>	<b>kunti.</b>
<b>mb</b>	mat	pol'nica-qit	ippi-sa-m	no:kur	irä-n-ti	kunti
<b>mp</b>	man	pal'nica-qin	ippi-si-m	no:kir	irä-n-ti	kunti
<b>ge</b>	I.NOM	hospital-LOC	lie-PST-1SG.O	three	month-GEN-3SG	during
<b>gr</b>	я.NOM	больница-LOC	лежать-PST-1SG.O	три	месяц-GEN-3SG	в.течение
<b>mc</b>	pers	n-n:case	v-v:tense-v:pn	num	n-n:case-n:poss	pp
<b>ps</b>	pers	n	v	num	n	pp
<b>SeR</b>	pro.h:Th	np:L			pp:Time	
<b>SyF</b>	pro.h:S		v:pred			
<b>IST</b>	giv-active	accs-gen				
<b>BOR</b>		RUS:cult				
<b>BOR-Phon</b>		Csub				
<b>BOR-Morph</b>		dir:infl				
<b>CS</b>						
<b>fr</b>	<i>Я лежал в больнице три месяца.</i>					
<b>fe</b>	<i>I've been in the hospital for three months.</i>					
<b>fg</b>	<i>Ich habe drei Monate im Krankenhaus verbracht.</i>					
<b>ltr</b>	<i>я в больнице лежал 3 месяца.</i>					
<b>nt</b>						
<b>nto</b>						

## 2.10.2. Transcription tiers

### 2.10.2.1. Main transcription tiers (tx and ts)

The main transcription tiers use the INEL Selkup transcription (see 2.6.2). 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, unintelligible fragments, false starts and non-speech sounds is described in a separate document:

[INELGuidelinesTranscriptionSpecialCases.pdf](#)

### 2.10.2.2. Source transcription tiers (st and stl)

For the texts from the written part of the archive, the source transcription tier (**st**) contains the original Cyrillic version of the text, as it appears in A. I. Kuzmina’s notebooks; the **stl** tier contains this original phonetic transcription, converted into Latin alphabet and simplified (see 2.6.2). For example, we omit stress marks and some of the diacritics above vowels, which provide additional phonetic information. In one text, namely KNS\_1966\_Markincha\_flk, published in [Tučkova, Helimski 2010], the published Latin transcription is provided in this tier. An example for these two tiers may be found in Figure 1.

For the texts from the sound part of the archive transcribed in fieldwork sessions, the **st** tier is empty, and the **stl** tier contains the original transcription made by researchers. For those audio texts transcribed by native speakers themselves, the **st** tier contains their original Cyrillic transcription, and the **stl** tier contains this original transcription converted into Latin alphabet.

### 2.10.3. Annotation tiers

#### 2.10.3.1. Reference (ref)

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

<b>ref</b>	KR_1969_RavensAndHares_flk.001 (001.001)	KR_1969_RavensAndHares_flk.002 (001.002)
------------	--	--

#### 2.10.3.2. Morpheme breaks (mb)

The morpheme breaks tier (**mb**) breaks words into segmentable morphs. Each word, according to the tier **tx**, appears in a separate cell. The morphs are represented in their surface form and are separated from each other by hyphens. Zero morphs are not represented in this tier. For an example see Figure 1.

#### 2.10.3.3. Morphemes (lexical) (mp)

The lexical morphemes tier (**mp**) shows the lexical representation of the morphs, both stems and affixes, which appear in the **mb** tier. It follows the FLEx lexicon which is, importantly, different for different dialects (Northern, Central, Southern except Ket, and Ket) (see 2.8).

For Northern dialects, this lexical representation is aligned with the dictionary [Kazakevich, Budyanskaya 2010] for stems and with the grammar sketch [Kuznecova et al. 1980] for affixes. For Central and Southern dialects, however, there is no single source one could rely upon, neither in lexicon nor in grammar. The cross-dialectal dictionary [Bykonina et al. 2005] lists numerous dialectal variants but does not posit any of them as a primary one. Therefore, the selection of the allomorph which appears in the **mp** tier is arbitrary for all dialects except Northern ones.

All morphemes within a word are separated by hyphens. Zero morphs are not represented in this tier. For an example see Figure 1.

#### 2.10.3.4. Gloss (ge, gr)

The gloss tiers (**ge**, **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 capital Latin letters and mostly according to the Leipzig Glossing Rules.<sup>7</sup> For the list of abbreviations used see [Appendix 3](#).

Glosses for all morphemes within a word are separated by hyphens. Non-overt morphemes are given in square brackets preceded by a dot (e.g. ".[3SG]").

If a morpheme 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").

Alternative meanings are separated by a slash (e.g. "ILL/LOC/EL").

Morphemes with unknown meaning are glossed with two percent signs ("%"). In case of uncertainty, the gloss is preceded with a single percent sign (e.g. "%whistle").

For an example see Figure 1.

#### 2.10.3.5. 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). Table 10 in [Appendix 4](#) shows the tags used for parts of speech and inflectional categories. For inflectional affixes the pattern x:a is used, where x stands for part of speech, to which an

<sup>7</sup> <https://www.eva.mpg.de/lingua/resources/glossing-rules.php>, last access: 01.11.2017.

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 an example see Figure 1.

### 2.10.3.6. 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 list of possible parts of speech can be found in [Appendix 4](#) (tags without ":", which are used for inflectional categories). Words with unknown part of speech are marked with two percent signs ("%").

For an example see Figure 1.

### 2.10.3.7. Syntactic function (SyF)

The annotation scheme used in the syntactic function tier was developed by Beáta Wagner-Nagy and Sándor Szevevényi [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. We also tag copulae in complex predicates. As well as semantic roles (see below), syntactic functions are only tagged in main clauses, with an exception for complement clauses in the form of direct speech. But subordinate clauses themselves are being tagged, the cells belonging to the subordinate clause are merged.

The full set of tags for SyF tier is provided in Table 2.

*Table 2. Tags for syntactic functions*

Tag	Description
<b>Main arguments</b>	
S	subject
O	direct object
<b>Predicate</b>	
v:pred	verbal predicate
n:pred	nominal predicate
adj:pred	attributive/adjectival predicate
pro:pred	pronominal predicate
ptcl:pred	particle predicate
cop	copula
<b>Subordinate clauses</b>	
s:comp	complement clause
s:rel	relative clause
s:temp	temporal clause
s:cond	conditional clause
s:adv	adverbial clause
s:purp	purpose clause

There are two other points that concern annotating both semantic roles and syntactic functions. First, we place the annotations (in corresponding tiers) on the head of the noun phrase, on the noun in the prepositional phrase and on the whole clause if it is a subordinate clause; for covert referents, the annotation is placed on the predicate. Second, we annotate the properties of the referent (if relevant) in both annotation layers. These properties are thus duplicated, for the more comfortable perception of tagging and also to make the search a bit easier. These properties are annotated before the main tag and are separated with a colon (<:>).

For each referent, we mark whether it is covert (<0>) or not (no special tag). In case the referent is covert (for example, it is a pro-drop subject), we indicate three possible values of the grammatical category "person": first person (<1>), second person (<2>) and third person (<3>). If the referent is overt, we annotate its form: whether it is a personal or demonstrative pronoun (<pro>), a noun phrase (<np>), postpositional phrase (<pp>) or an adverbial phrase (<adv>). For both overt and covert referents we annotate, whether they are human (<h>) or non-human (no tag). Tags for different properties of a referent are separated with a dot (<.>).

Table 3. 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
v	verb

Illustration of referent annotation may be found in the examples below and in section 2.10.3.8, since it is part of both semantic roles and syntactic functions annotations.

Here are some examples for tagging syntactic functions:

Figure 2.

ref	SAIAn_1965_Soldatka_nar.026 (003.004)				
st	мѐ 'клубмын 'орса сѐм еџа.				
tx	Me:	klupmin	orsa	som	ε:џа.
mb	me:	klup-mɪn	or-sa	som	ε:-џа
ge	we.PL.GEN	club.[NOM]-1PL	force-INSTR	good	be-CO.[3SG.S]
ps	pers	n	n	adj	v
SyF		np:S		adj:pred	cop
fe	<i>Our club is very good.</i>				

Figure 3.

ref	KPG_1969_MyFamilyAndMyVillage_nar.006 (001.006)					
st	м̄ат мѐле 'туртукула п̄џа прикащ'има 'т̄сона.					
tx	M̄a:t	me:l̄ä	tu:rtukula	pu:la	prikaš'ima	t̄sona.
mb	m̄a:t	me:-l̄ä	tu:r-tu-ku-la	pu:la	prikaš'ima	t̄ü-so-na
ge	house.[NOM]	make-CVB	be.over-TR-HAB-CVB	after	saleswoman.[NOM]	come-PST-3SG.S
ps	n	adv	adv	ptcl	n	v
SyF	s:temp				np.h:S	v:pred
fe	<i>After they have finished to build the house, a saleswoman arrived.</i>					

In case of conjunction of predicates we duplicate the information about the subject in the cell of the second predicate:

Figure 4.

ref	SAIAn_1965_Soldatka_nar.028 (003.006)							
st	'Ви́ча 'коныша 'лапкoнды и 'таттысыды 'н'әјеп и 'ч̣̣̣п[̣̣̣]ы.							
tx	Vič'a	qoniša	lapkonti	i	ta:ttisiti	n'äjep	i	č'ɔ:pi.
mb	Vič'a	qon-i-ša	lapko-nti	i	ta:tti-si-ti	n'äj-e-p	i	č'ɔ:pi
ge	Vitya.[NOM]	leave-EP-PST.[3SG.S]	shop-ILL	and	bring-PST-3SG.O	bread-EP-ACC	and	tobacco.[NOM]
ps	nprop	v	n	conj	v	n	conj	n
SyF	np:h:S	v:pred			0.3.h:S v:pred	np:O		np:O
fe	<i>Vitya went to the shop and brought some bread and tobacco.</i>							

We tag null subjects and null objects in the cell of the predicate:

Figure 5.

ref	KPG_1969_MyFamilyAndMyVillage_nar.019 (001.019)				
st	'әсыты тү́лла 'пү́ла 'и́лла 'ме'сoтын.				
tx	Әsiti	tülla	pu:la	illa	me:sɔ:tin.
mb	әsi-ti	tü-lla	pu:la	illa	me:sɔ:-tin
ge	father.[NOM]-3SG	come-CVB	after	down	make-PST-3PL
ps	n	adv	ptcl	preverb	v
SyF	s:temp				0.3.h:S 0.3.h:O v:pred
fe	<i>After his father had come they buried him.</i>				

Figure 6.

ref (ref)	KPG_1969_Bread_nar.005 (001.005)		
st (st)	'штә́лле шoүортä 'тoттäдъ.		
tx (tx)	Štäl'l'e	šo:qortä	tottätä.
mb (mb)	štäl'l'e	šo:qor-tä	tott-ätä
ge (ge)	then	oven-ILL	put-IMP.2SG.O
ps (ps)	adv	n	v
SyF (SyF)			0.2.h:S 0.3:O v:pred
fe (fe)	<i>Then put it into the oven.</i>		

We do not tag subjects and objects in case they are expressed by complement clauses (see Figure 7 and Figure 8). The only exception is when the object is expressed by a single predicate without any dependent words, in which case it is marked as v:O (see Figure 9).

Figure 7.

ref	SAIAn_1965_Soldatka_nar.020 (002.018)				
st	'täпынык 'үчико сав'хоскын 'сәтым 'әйса.				
tx	Täpinik	u:č'i:qo	savhosqin	sätim	ε:jsa.
mb	täp-i-nik	u:č'i-qo	savhos-qin	sätim	ε:j-sa
ge	(s)he-EP-ALL	work-INF	sovkhos-LOC	difficult	be-PST.[3SG.S]
ps	pers	v	n	adj	v
SyF		s:compl		adj:pred	cop
fe	<i>It was hard for her to work in sovkhos.</i>				

Figure 8.

ref	SAIAn_1965_Soldatka_nar.014 (002.012)					
st	чап 'түс[э]а 'моткынты. 'кōсыты' илл[а] 'олджимба.					
tx	Čap	tüsa	mә:tqinti,	qositi	illə	ә:l'č'impa.
mb	č'ap	tü-sa	mә:t-qin-ti	qo-si-ti	illə	ә:l'č'i-mpa
ge	hardly	come-PST.[3SG.S]	house-ILL-3SG	sight-PST-3SG.O	down	fall.down-PST.NAR.[3SG.S]
ps	conj	v	n	v	preverb	v
SyF	s:temp			0.3.h:S v:pred	s:compl	
fe	<i>As she came to her house, she saw, that it had been destroyed.</i>					

Figure 9.

ref	KuLP_1976_Hospital_nar.002 (001.002)	
st	'илоко 'кыкаң.	
tx	iloqo	kikañ.
mb	ilo-qo	kika-ñ
ge	live-INF	want-1SG.S
ps	v	v
SyF	v:O	0.1.h:S v:pred
fe	<i>I want to live.</i>	

For embedded subordinate clauses, only the external clause is tagged:

Figure 10.

ref	KIA_1965_Petro_transl.006 (001.006)			
st	ны'най 'щ'ентың[к] пўкылтыки'жоим'пелым`пат,			
tx	Ni:n	aj	šentiñ	püqiltikkijoimpelimpat
mb	ni:n	aj	šenti-ñ	pü-qil-ti-kki-j-oim-pe-li-mpa-t
ge	then	again	new-ADVZ	touch-MULO-IPFV-HAB-INF-be.going.to-DUR-INCH-PST.NAR-3SG.O
ps	adv	adv	adv	v
SyF				0.3.h:S v:pred

st	на н'ўты 'мōтам ка'талпытый 'пōтпыкын`тоу[к]а.				
tx	na	n'у:tə	mә:tam	qatalpitij	pōtpiqinto:qa.
mb	na	n'у:tə	mә:ta-m	qat-al-pi-tij	pōt-pi-qinto:qa
ge	this	hay.[NOM]	door-ACC	hit-MOM-DUR-PTCP.PRS	warm-DUR-SUP.2/3SG
ps	dem	n	n	ptcp	v
SyF			np:O	s:rel	
fe	<i>Then he started to touch the door padded with hay for warmth isolation.</i>				

### 2.10.3.8. Semantic roles (SeR)

The annotation of semantic (thematic) roles is given in tier labelled **SeR**. It is based on GRAID principles [Haig & Schnell 2014] with some further developments by Beáta Wagner-Nagy and Sándor Szeverényi [Wagner-Nagy & Szeverényi 2018: 21ff.], further adapted for the current project.

The full set of tags for semantic roles is listed in Table 4.

Table 4. Tags for semantic roles

Tag	Description
<b>A</b>	Agent: initiator (with volition) of the action, the participant is causing the action or it is responsible for something happening.
<b>B</b>	Beneficiary: entity for whose benefit the action is being performed.
<b>Com</b>	Comitative: entity that conveys a participant of an action.
<b>Cau</b>	Cause: entity that causes an event.
<b>E</b>	Experiencer: entity that experiences the action, it does not have a control of the action or state (first argument of the verbs of emotion, volition, cognition, perception).
<b>G</b>	Goal: location or entity in the direction of which something moves.
<b>Ins</b>	Instrument: medium by which the action or event is performed.
<b>L</b>	Locative: locative argument of verb, place in which something is situated (states location)
<b>P</b>	Patient: entity, which undergoes physical changes, is created or destroyed.
<b>Path</b>	Path: entity or location along or through which the event takes place.
<b>Poss</b>	Possessor: entity which owns something.
<b>R</b>	Recipient: animate recipient of transfer or the addressee of verb of speech.
<b>So</b>	Source: place of origin or original owner in a transfer.
<b>Th</b>	Theme: entity which does not undergo physical changes, but is affected otherwise by an action (change of location or possession: object of give; subject of walk); entity whose location is specified; the content of mental verbs and verba dicendi (think, say etc.); second argument of verbs like <i>see, hear, feel</i> .
<b>Time</b>	Time: time point or an interval of time.

This list does not pretend to cover all possible semantic functions, since we do not aim to tag every noun phrase in the text, we tag the most frequent ones. For example, we do not tag semantic roles for depictives or translatives.

As well as for syntactic functions, we do not annotate semantic roles inside subordinate clauses. Referent annotation follows the same rules as in the SyF tier (see section 2.10.3.7).

Here is an example of SeR-tagging:

Figure 11.

ref	SMI_1965_IWasBornInChaselka_nar.002 (001.002)		
st	'нымтӓ 'илесак 'мелды.		
tx	Nımtä	ilesak	me:lti.
mb	nımtä	ile-sa-k	me:lti
ge	here	live-PST-1SG.S	all.the.time
ps	adv	v	adv
SeR	adv:L	0.1.h:Th	adv:Time
fe	<i>I lived here all the time.</i>		

#### 2.10.3.9. 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 & Szevéryni (2016: 20ff.) 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 [ $\pm$ discourse-old] and [ $\pm$ hearer-old]:

Table 5. Parameters for determining information status

	<b>+discourse-old</b>	<b>- discourse-old</b>
<b>+hearer-old</b>	given	accessible
<b>- hearer-old</b>	---	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 6. Basic tags for annotating information status

<b>Tag</b>	<b>Description</b>
<b>Given referents</b>	
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)
<b>Accessible referents</b>	
accs-sit	referent accessible through the situation (e.g. having breakfast: “Give me <u>the butter</u> , please.”)
accs-aggr	referent accessible through the aggregation of other referents (e.g. “ <i>Unce upon a time, a king had a wife and two children. <u>They</u> lived happily.</i> ”)
accs-inf	referent accessible through inference, e.g. part-whole relations (e.g. “ <i>We had a turkey for thanksgiving. I ate its <u>wings</u>.</i> ”)
accs-gen	referent accessible through general knowledge (e.g. “ <i>The <u>president of the U.S.</u> travelled to Cuba.</i> ”)
<b>New referents</b>	
new	new referent

Here is an example of tagging the information structure:

Figure 12.

ref	KNK_1965_BearAndHare_flk.012 (001.012)			
st	нөма 'у́кыт 'кāлымпа 'сүрыки `jat			
tx	N'oma	u:kit	qälimpa,	su:rikijat
mb	n'oma	u:ki-t	qäli-mpa	su:ri-k-ija-t
ge	hare.[NOM]	front.part-ADVZ	run-PST.NAR.[3SG.S]	wild.animal-DIM-child-PL.[NOM]
ps	n	adv	v	n
IST	giv-active			giv-inactive

st	'ныркы `молле 'мāчонте ку `ррәлнөты[ә]т.		
tx	nirki:m:llä	mač'onte	kurəlnɔ:tit.
mb	nirki-mɔ:l-lä	mač'o-nte	kur-əl-nɔ:-tit
ge	be.frightened-DECAUS-CVB	forest-ILL	run-MOM-CO-3PL
ps	adv	n	v
IST		accs-gen	
fe	<i>The hare ran in front, the baby wild animals were frightened and ran away into the forest.</i>		

In this example the hare was mentioned in a previous sentence and it is thus “giv-active”, the baby animals were mentioned earlier, but not in the previous sentence – they are “giv-inactive”, and the forest, though not mentioned before, is a well-known place for animals to live, and thus it gets an “accs-gen” status.

For non-overt referents, the prefix <0.> is added to the corresponding tag (e.g. <0.giv-active> for a zero/covert given and active referent) placed at the predicate of the clause.

Another extension of the tag system is related to direct speech. As is widely known, direct speech tends to change to perspective of both the hearer and the speaker which has consequences for the discourse status of referents as well. In the present corpus the information status of referents in direct speech only reflects the level of the macro-discourse, i.e. the whole communication. However, in order to be aware of possible changes of perspective, the tag <-Q> is added, if a referent occurs in direct speech (e.g. <accs-gen-Q>, i.e. a referent, accessible through general knowledge in direct speech), as it is done in the Nganasan Spoken Language Corpus (NSLC) [Brykina et al. 2018] according to Wagner-Nagy et al. [2018: 30]. Furthermore, so-called utterance predicates are tagged with <quot-sp>.

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

A detailed information about these tiers is provided in a separate guidelines document:

[INELGuidelinesBorrowingCodeSwitching.pdf](#).

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 Selkup, Russian (RUS) borrowings are tagged. For the type of borrowing the following tags are used:

Table 7. 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

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

Table 8. Annotation panel for phonological processes in borrowings

Tag	Description
<b>Deletions</b>	
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)
<b>Insertions</b>	
inVins	initial vowel insertion
medVins	medial vowel insertion
finVins	final vowel insertion
<b>Substitutions</b>	
Csub	consonant substitution
Vsub	vowel substitution
<b>Other</b>	
lenition	lenition (weakening)
fortition	fortition (strengthening)

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

Table 9. Tags for annotating morphological processes in borrowings

Tag	Description
<b>Adaptation strategies</b>	
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)
<b>Further inflection (in the matrix language)</b>	
:bare	no inflection
:infl	further inflection

Here is an example of tagging a word borrowed from Russian:

Figure 13.

ref	KuLP_1976_Hospital_nar.009 (001.009)				
st	ман пол''ницауын 'иппысаң, 'чөзың асысам.				
tx	Man	pol'nicaqin	ippisaŋ,	č'ösinq	esisam.
mb	man	pol'nica-qin	ippi-sa-ŋ	č'ös-i-ŋ	esi-sa-m
ge	I.NOM	hospital-LOC	lie-PST-1SG.S	fat-EP-ADVZ	become-PST-1SG.O
ps	pers	n	v	n	v
BOR		RUS:cult			
BOR-Phon		Csub			
BOR-Morph		dir:infl			
fe	<i>I have been in the hospital, I got fat.</i>				

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 according to the scheme <LANGUAGE:TYPE>. The language is mostly Russian (RUS).

In the same tier, grammatical calques from Russian are marked; this does not interfere with code-switching tags since calques appear in Selkup fragments.

The tag set for the CS tier is the following:

Table 10. Tags for annotating code-switching

Tag	Description
<b>Sentence-external code-switching</b>	
:ext	languages change at sentence (clause, utterance) borders
<b>Sentence-internal code-switching</b>	
: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
<b>Calque</b>	
:calq	calque

Here is an example of code switching:

Figure 14.

ref	KAI_1965_OldManWithLittleMind1_flk.049							
st	'коты яйцаты то много, чинкып 'йса `мёна ('ёҥтоко) ['еҥты чоты].							
tx	Qɔ:ti	jajcatɨ	to	mного,	č'ɨŋkip	i:sam	ena	(eŋto:qo)...
mb	qɔ:ti	jajca-tɨ	to	mного	č'ɨŋki-p	i:-sa-m	ena	eŋ-to:qo
ge	probably	egg.[NOM]-3SG	EMPH2	many	swan-ACC	take-PST-1SG.O	CONJ	egg-TRL.3SG
ps	ptcl	n	ptcl	quant	n	v	ptcl	n
BOR		RUS:core	RUS:disc	RUS:core				
CS			RUS:int.alt					
fe	"I could probably get a lot of eggs, I would take a swan for the eggs".							

### 2.10.3.11. Free translation (fe, fr, fg)

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

[INELGuidelinesTranslationsandComments.pdf](#)

### 2.10.3.12. Literal Russian translation (ltr)

For texts from the written part of the archive the literal Russian translation tier (ltr) contains the original Russian translation of the sentence in question made by A. I. Kuzmina. Sometimes it is a word-by-word translation, sometimes it is a literary translation, sometimes it is something in between. For texts translated from Russian originals the source text, which is being translated, is provided here. For texts from the sound part of the archive this tier contains original translation made by native speakers while transcribing.

### 2.10.3.13. Notes (nt, nto)

The Notes tier (nt) 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 section 2.6.8) in square brackets, followed by a colon).

For texts from the written part of the archive the original notes tier (nto) contains notes from the notebooks of A. I. Kuzmina. These can be either comments to the texts or translation of separate words. Usually these comments are made by A. I. Kuzmina herself. In the text KNS\_1966\_Markincha\_flk, which is a single case in the archive of native speaker transcribing himself, the notes are made by the author. For texts from the sound part of the archive this tier contains notes made by native speakers while transcribing.

## References

- Arkhipov A. V., Däbritz C. L. 2018. Hamburg corpora for indigenous Northern Eurasian languages // *Tomsk Journal of Linguistics and Anthropology*. Issue 3 (21). P. 9–18. URL: [https://ling.tspu.edu.ru/en/archive.html?year=2018&issue=3&article\\_id=7130](https://ling.tspu.edu.ru/en/archive.html?year=2018&issue=3&article_id=7130)
- Brykina Maria, Gusev Valentin, Szeverényi Sandor, Wagner-Nagy Beáta. 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>
- Bykonja V. V., Kuznecova N. G., Maksimova M. P. 2005. *Селькупско-русский диалектный словарь [Selkup-Russian dialectal dictionary]*. Tomsk: Tomsk State Pedagogical University.
- Glushkov S.V., Bajdak A.V., Maksimova N.P. 2013. Диалекты селькупского языка [Dialects of Selkup]. In Tuchkova et al. 2013, p. 49–63.
- Götze, Michael et al. 2007. Information structure. In Dipper, S., Götze, M. and S. Skopeteas (eds), *Information Structure in Cross-Linguistic Corpora*. Interdisciplinary Studies on Information Structure 07. P. 147–187. URL: [https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/deliver/index/docId/2036/file/Kapitel6\\_07.pdf](https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/deliver/index/docId/2036/file/Kapitel6_07.pdf) [Accessed: 02.11.2017].
- Haig, Geoffrey; Schnell, Stefan. 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](https://www.uni-bamberg.de/fileadmin/aspra/Publications/GRAID7.0_manual.pdf) [Accessed: 01.11.2017].
- Helimski E.A. 1998. Selkup. In Abondolo, Daniel (ed.), *The Uralic languages*. London. P. 548–579.
- Kuznecova A. I., Helimsky E. A., Grushkina E. V. 1980. *Очерки по селькупскому языку: Тазовский диалект [Essays on Selkup: Taz dialect]*. Moscow: Moscow University Press.
- Kazakevich O. A., Budyanskaya E. M. 2010. *Диалектологический словарь селькупского языка (северное наречие) [Dialectological dictionary of Selkup (Northern dialect)]*. Moscow: Basko.
- Tuchkova N.A., Glushkov S.V., Golovnev A.V., Kosheleva E.Yu. et al. 2013. *Селькупы: Очерки традиционной культуры и селькупского языка [The Selkup: Essays on traditional culture and the Selkup language]*. Tomsk.
- Tučkova N.A., Helimski E.A. 2010. Über die selkupischen Sprachmaterialien von Angelina I. Kuzmina (*О материалах А. И. Кузьминой по селькупскому языку*). *Hamburger Sibirische und Finno-Ugrische Materialien*, Bd. 5. Hamburg.
- VPN 2010 = Vserossijskaya perepis` naseleniya 2010. Tom 4. Nacional`ny`j sostav i vladenie yazy`kami. URL: [http://www.gks.ru/free\\_doc/new\\_site/perepis2010/croc/Documents/Vol4/pub-04-05.pdf](http://www.gks.ru/free_doc/new_site/perepis2010/croc/Documents/Vol4/pub-04-05.pdf). [Accessed: 06.11.2017].
- Wagner-Nagy, Beáta; Szeverényi, Sándor; Gusev, Valentin. 2018. User’s Guide to Nganasan Spoken Language Corpus. // *Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology*. Vol. 1, no. 1. P. 1–45. URL: <http://www.iskolakultura.hu/index.php/wpcl/article/view/10611/10503>. [Accessed: 17.06.2018].
- Published texts from Angelina Kuzmina’s Selkup archive
- Dul`zon A.P. 1966. “Ketskie skazki”. Tomsk. P. 128–155.
- Kuzmina A.I. 1967. “Dialektologicheskie materialy po sel`kupskomu jazy`ku” // *Issledovanija po jazy`ku i folkloru*. Issue II. Novosibirsk. P. 267–329.
- Kuzmina A.I. 1974. “Grammatika sel`kupskogo jazy`ka”, Novosibirsk.
- Kuzmina A.I. 1977. “K etimologii nazvanij mesyacev, storon sveta, zvyozd i sozvezdij v sel`kupskom jazy`ke” // *Yazy`ki i toponimiya*. Issue IV. Tomsk. P. 71–85.
- Tutschkova N.A., Wagner-Nagy, B. 2015. Семи богов мудростью обладающий Итте...» Тексты с героем Итя в селькупском фольклоре. [Texte über Itte, der über die sieben Götter der Weisheit verfügt]. Tomsk. P. 65–76, 78–81, 89–92, 100–105, 141–146, 183–185, 194–197, 206–214, 215–220, 223–242, 256–264.

## Appendix 1. INEL transcription for Northern Selkup dialects

Table 1. INEL Selkup transcription characters

INEL Transcription	Description	Unicode Character Name
<b>Vowels</b>		
a	low central unrounded vowel	LATIN SMALL LETTER A (U+0061)
a:	low central unrounded long vowel	LATIN SMALL LETTER A + MODIFIER LETTER TRIANGULAR COLON (U+02D0)
ä	low front unrounded vowel	LATIN SMALL LETTER A WITH DIAERESIS (U+00E4)
ä:	low front unrounded long vowel	LATIN SMALL LETTER A WITH DIAERESIS (U+00E4) + MODIFIER LETTER TRIANGULAR COLON (U+02D0)
e	mid front tense unrounded vowel	LATIN SMALL LETTER E (U+0065)
e:	mid front tense unrounded long vowel	LATIN SMALL LETTER E (U+0065) + MODIFIER LETTER TRIANGULAR COLON (U+02D0)
ə	mid central unrounded vowel	LATIN SMALL LETTER SCHWA (U+0259)
ə:	mid central unrounded long vowel	LATIN SMALL LETTER SCHWA (U+0259) + MODIFIER LETTER TRIANGULAR COLON (U+02D0)
ɛ	mid front lax unrounded vowel	LATIN SMALL LETTER OPEN E (U+025B)
ɛ:	mid front lax unrounded long vowel	LATIN SMALL LETTER OPEN E (U+025B) + MODIFIER LETTER TRIANGULAR COLON (U+02D0)
i	high front tense unrounded vowel	LATIN SMALL LETTER I (U+0069)
i:	high front tense unrounded long vowel	LATIN SMALL LETTER I (U+0069) + MODIFIER LETTER TRIANGULAR COLON (U+02D0)
ɪ	high front lax unrounded vowel	LATIN SMALL LETTER DOTLESS I (U+0131)
ɪ:	high front lax unrounded long vowel	LATIN SMALL LETTER DOTLESS I (U+0131) + MODIFIER LETTER TRIANGULAR COLON (U+02D0)
ɨ	high central unrounded vowel	LATIN SMALL LETTER I WITH STROKE (U+0268)
ɨ:	high central unrounded long vowel	LATIN SMALL LETTER I WITH STROKE (U+0268) + MODIFIER LETTER TRIANGULAR COLON (U+02D0)
o	mid back rounded tense vowel	LATIN SMALL LETTER O (U+006F)
o:	mid back rounded tense long vowel	LATIN SMALL LETTER O (U+006F) + MODIFIER LETTER TRIANGULAR COLON (U+02D0)
ö	mid front rounded vowel	LATIN SMALL LETTER O WITH DIAERESIS (U+00F6)
ö:	mid front rounded long vowel	LATIN SMALL LETTER O WITH DIAERESIS (U+00F6) + MODIFIER LETTER TRIANGULAR COLON (U+02D0)
ɔ:	mid back rounded lax long vowel	LATIN SMALL LETTER OPEN O (U+0254) + MODIFIER LETTER TRIANGULAR COLON (U+02D0)
u	high back rounded vowel	LATIN SMALL LETTER U (U+0075)
u:	high back rounded long vowel	LATIN SMALL LETTER U (U+0075) + MODIFIER LETTER TRIANGULAR COLON (U+02D0)
ü	high front rounded vowel	LATIN SMALL LETTER U WITH DIAERESIS (U+00FC)
ü:	high front rounded long vowel	LATIN SMALL LETTER U WITH DIAERESIS (U+00FC) + MODIFIER LETTER TRIANGULAR COLON (U+02D0)
<b>Consonants</b>		
č'	alveolo-palatal affricate	LATIN SMALL LETTER C WITH CARON (U+010D) + MODIFIER LETTER APOSTROPHE (U+02BC)

<b>INEL Transcription</b>	<b>Description</b>	<b>Unicode Character Name</b>
j	palatal approximant	LATIN SMALL LETTER J (U+006A)
k	voiceless velar stop	LATIN SMALL LETTER K (U+006B)
l	alveolar lateral approximant	LATIN SMALL LETTER L (U+006C)
l'	alveolar palatalized lateral approximant	LATIN SMALL LETTER L (U+006C) + MODIFIER LETTER APOSTROPHE (U+02BC)
m	bilabial nasal	LATIN SMALL LETTER M (U+006D)
n	alveolar nasal	LATIN SMALL LETTER N (U+006E)
n'	palatalized nasal	LATIN SMALL LETTER N (U+006E) + MODIFIER LETTER APOSTROPHE (U+02BC)
ŋ	velar nasal	LATIN SMALL LETTER ENG (U+014B)
p	voiceless bilabial stop	LATIN SMALL LETTER P (U+0070)
q	voiceless uvular plosive	LATIN SMALL LETTER Q (U+0071)
r	voiced alveolar trill	LATIN SMALL LETTER R (U+0072)
s	voiceless alveolar fricative	LATIN SMALL LETTER S (U+0073)
š	voiceless postalveolar fricative	LATIN SMALL LETTER S WITH CARON (U+0161)
t	voiceless alveolar stop	LATIN SMALL LETTER T (U+0074)
w	labiovelar approximant	LATIN SMALL LETTER W (U+0077)

*Table 2. Phonemes which occur only in loanwords (mostly from Russian)*

<b>INEL Transcription</b>	<b>Description</b>	<b>Unicode Character Name</b>
b	voiced bilabial stop	LATIN SMALL LETTER B (U+0062)
c	voiceless alveolar affricate	LATIN SMALL LETTER C (U+0063)
d	voiced alveolar stop	LATIN SMALL LETTER D (U+0064)
f	voiceless labiodental fricative	LATIN SMALL LETTER F (U+0066)
g	voiced velar stop	LATIN SMALL LETTER G (U+0067)
v	voiced labiovelar fricative	LATIN SMALL LETTER V (U+0076)
x	voiceless velar fricative	LATIN SMALL LETTER X (U+0078)
z	voiced alveolar fricative	LATIN SMALL LETTER Z (U+007A)
ž	voiced postalveolar fricative	LATIN SMALL LETTER Z WITH CARON (U+017E)

As the transcription system of the INEL project has some differences in graphic representation of the Northern Selkup phonemes with the existing grammars and dictionaries, here is the table which illustrates correspondences in three transcription systems of the Northern Selkup (INEL Project, [Kuznecova et al. 1980], [Kazakevich, Budyanskaya 2010]).

Table 3. Comparative transcription chart

INEL Transcription	Example	Transcription in [Kuznecova et al. 1980]	Example	Transcription in [Kazakevich, Budyanskaya 2010]	Example	English gloss
A, a	anti	A, a	anty	A, a, Я, я*	анты	'boat'
A:, a:	ma:tiqo	Ā, ā	mātyqo	Ā, ā, Я, я*	мāтықо	'cut'
Ä, ä	täpäŋ	Ä, ä	täpäŋ	Ä, ä	тāпāң	'squirrel'
Ä:, ä:	č'ä:ŋkɨ	Ä, ä	cāŋky	Ä, ä	чāңкы	'no(t)'
Č, č'	č'u	C, c	cu	Ч, ч	чу	'earth'
E, e	keti	E, e	kety	E, e	кеты	'intestine'
E:, e:	č'e:li	Ē, ē	cēly	Ē, ē	чēлы	'day'
Ə, ə	əsi	Ə, ə	əsy	Ə, ə	əсы	'father'
Ə:, ə:	ə:ti	Ē, ē	ēty	Ē, ē	ēты	'word'
ɛ, ɛ	esiqo	ɛ, ɛ	esyqo	Э, э	эсықо	'become'
ɛ:, ɛ:	ɛ:qo	Ē, ē	ēqo	Ē, ē	ēқо	'be'
İ, i	ira	I, i	ira	И, и	ира	'old man'
İ:, i:	i:ja	Ī, ī	īja	Й, й	йя	'son'
l, l	illä	l, l	illä	l, l	ллā	'down'
l:, l:	li:pɨ	Ī, ī	līpy	Ī, ī	лīпы	'piece'
t, t	inti	Y, y	ynty	Ы, ы	ынты	'bow'
t:, t:	i:tiqo	Ȳ, ȳ	ȳtyqo	Ы, ы	ытықо	'hang'
J, j	qaj	J, j	qaj	Й, й**	қай	'what'
K, k	kəm	K, k	kəm	К, к	кәм	'blood'
L, l	loqa	L, l	loqa	Л, л	лоқа	'fox'
L', l'	l'aqa	L', l'	l'aqa	Л, л, ль***	ляқа	'friend'
M, m	mač'i	M, m	macy	М, м	мачы	'forest'
N, n	nom	N, n	nom	Н, н	ном	'sky'
N', n'	n'en'n'a	Ñ, ñ	ñeñña	Н, н, нь***	неньня	'sister'
ŋ	taŋɨ	ŋ	taŋy	Ң	таңы	'summer'
O, o	oli	O, o	oly	О, о, Ё, ё*	олы	'head'
O:, o:	lo:sɨ	Ō, ō	lōsy	Ō, ō, Ё, ё*	лōсы	'devil'
ɔ:, ɔ:	mɔ:t	Ǫ, ǫ	mǫt	Ǫ, ǫ	мǫт	'tent'
Ö, ö	pöt	Ö, ö	pöt	Ö, ö	пöт	'warm'
Ö:, ö:	tö:ka	Ö, ö	tōka	Ö, ö	тōка	'goose'
P, p	pi	P, p	pi	П, п	пи	'night'
Q, q	qup	Q, q	qup	Қ, қ	қуп	'man'
R, r	ruš	R, r	ruš	Р, р	руш	'Russian'
S, s	soma	S, s	soma	С, с	сома	'good'
Š, š	ši:pa	Š, š	šipa	Ш, ш	шйпа	'duck'
T, t	tama	T, t	tama	Т, т	тама	'mouse'
U, u	ukkir	U, u	ukkyr	У, у, Ю, ю*	уккыр	'one'

INEL Transcription	Example	Transcription in [Kuznecova et al. 1980]	Example	Transcription in [Kazakevich, Budyanskaya 2010]	Example	English gloss
U:, u:	tu:	Ū, ū	tū	Ū, ū, Ю, ю*	тū	'feather'
Ü, ü	ür	Ü, ü	ür	Û, ù	Ûр	'fat'
Û:, ù:	ü:tɨ	Ū, ū	ūty	Û, ù	Ûты	'evening'
W, w	wəɾqɨ	W, w	wəɾqy	В, в	вəɾқы	'big'

In the dictionary [Kazakevich, Budyanskaya 2010] some specific rules apply to these Cyrillic symbols:

\* я, ю, ё are used word-initially for sequences of [j] + a, o, y (e.g. яннə) and after the palatalized l', n' (ляжа, нюр);

\*\* ū is used word-initially for [j] when followed by a vowel other than a, o, y (e.g. ūəвал);

\*\*\* нь, ль are used before ə, ē (e.g. ньəмпы, льəсқымпықо).

## Appendix 2. Conversion into INEL transcription

### Northern Selkup dialects: Angelina Kuzmina's transcription

*Table 4. Automatic conversion of A. I. Kuzmina's transcription into INEL transcription for Northern Selkup dialects*

<b>Kuzmina's transcription</b>	<b>INEL transcription</b>
<b>Vowels<sup>8</sup></b>	
а	a
ä	ä
е	e
и	i
о	o
ö	ö
у	u
ÿ	ü
ъ	ə (U+0259)
ы	ɨ (U+0268)
э	ɛ (U+025B)
<b>Consonants</b>	
ʏ	q
к	q
л	l'
б	p
в	v
г	g
д	d
ж	ž (U+017E)
з	z
й	j
к	k
л	l
м	m
н	n
ң	ŋ (U+014B)
п	p
р	r
с	s
т	t
т'	č (U+010D)
ф	f
х	h
х̣	q
ц	c
ч	č (U+010D)
ш	š (U+0161)
ш'	š (U+0161)
щ	š (U+0161)

<sup>8</sup> No automatic conversion was made for uncommon vowels ё, ю, я.

Latin symbols left unchanged	
ε	ε (U+025B)
ə	ə (U+0259)
j	j
w	w
Diacritics	
˘ (U+0304)	: (U+02D0)
ь	' (U+02BC)
Ignored symbols	
° (U+030A)	
◦ (U+0325)	
˜ (U+0303)	
˙ (U+0351)	
˚ (U+0328)	
˘˘ (U+2032)	
˘˘˘ (U+2035)	

Table 5. Changes made manually to A. I. Kuzmina's transcription for the texts from the Northern dialect.

Original Kuzmina's transcription	INEL transcription	Context	Examples
<b>Vowels</b>			
short / long vowels	long / short vowels	according to the dictionary* and morphonological rules**	ȳгон - ȳko:n ытысыты - i:tisiti чон'докын - č'onto:qin
и	i	according to the dictionary	нѣк - ni:k чирымпа - č'irimpa
ы	i	stem DU	ыннă - innä тымн'ăкынтыкине - timn'ăqintikine
о	ɔ:	stem, before verbal endings  stem	метомын - me:tɔ:mɪn илымпōтын - ilɪmpɔ:tɪt илымпоки - ilɪmpɔ:qi  чоты - čɔ:ti ноқыр - no:kɪr мотыр - mo:tɪr
у	ü	after ш('), щ, ч	ш'ун'анты - šün'n'anti чушимпа - č'üšimpa
ÿ	u	after н'	н'ÿтысă - n'u:tisă
ö	o, ɔ:	after н'	н'öтыңыт - n'o:tɪŋt има'н'öтыт - iman'n'o:tɪt
e, ä	ε	FUT  PFV  stem	чотентал - čɔ:tɛntal пинтăртăнтал - pintărtɛntal  контыш'ента - qontiššɛ:nta  тăнырна - tɛnɪrna ăсымпа - ɛsimpa
ę, e	ε:	verb «ε:qo»	ęja - ɛ:ja ęппынтотыт - ɛ:ppɪnto:tɪt

е, ә	ä	stem  CVB	ылле - illä албекыт - alpəqıt емә - ε:mä мүттыл'е - müttil'ä
ä	e	stem	н'әннәнты - n'ennänti н'әнты - n'enti м'әш'ак - me:šak
ѳ, ә	ö	stem	тѳса - tōsa кәнты - qōnti
ö	ә	stem	төпыт - təpıt
а, ä	ɔ	before 1/3PL  before DU stem	месамын - me:sɔ:mɪn йчирпатыт - ič'č'irpɔ:tıt антәки - antɔ:qi чәппаймпаты - č'ɔ:ppai:mpati
ъ	i	stem ending  III	кѳттѳ - qə:ttı поркѳ - porqı мүтонтѳ - mütontı мөттѳ - mɔ:ttı
<b>Consonants</b>			
voiced consonants	voiceless consonants	according to phonematic principle	лѳзы - lo:sı анды - antı
к, k	q	according to the dictionary	кйш'кат - qıšqat қарын - qarın
k	k	according to the dictionary	кѳтсанмы - kətsanmı кәтысыт - kätısıt
ү	q	everywhere except PRS	әтаұты - ɔ:taqıtı (but: еұанты - ε:ɯanti)
x	k	word ending	чѳташ'их - č'ɔ:täšik
т'	č'	stem	т'елакса - č'e:laksa
л	l'	ADJZ	ырал - ira'l
н	n'	before i	мѳкыңыт - me:qıŋ'ıt
ш', щ'	š	US IMP.2SG.S stem	пис'еш'па - piššpa чѳташ'их - č'ɔ:täšik омбашим - ompašim wəш'импа - wəšimpa
с'	s	CON.REC	тымн'ас'ыт - timn'asıt
нк	ŋ, ŋk	according to the dictionary	нынқа - niŋa чәнқа - č'ä:ŋka
г	ŋ	before sonorants	мѳтыглымыт - mɔ:tıŋlɪmıt
н	ŋ	IMP	амырңылыт - amırŋılıt
тч	č'č'	according to the dictionary	сатчи`мөтын - sač'č'imɔ:ttin
any consonant	double consonant	in verbs before PFV	кәтәмыт - qätte:mıt
н*	nn	stem  morphemic boundary	ина - innä коңә - konnä кәңә - qännä

нн*	n	stem ending ILL PRS	пўннон - pünin әмыннты - әmĩntĩ пактынна - paktĩna
м*	mm	PST.NAR according to morphological rules	ил'имынтотыт - ilĩmmĩnto:tit
мм*	m	stem ending	куммыл' - qumĩl'
л, л*	ll	stem morphemic boundary	йла - ilä оркылде - orqĩllä
лл, ll*	l	stem CVB RES	илла - ilä тантылла - tantĩlä тўлыл'лак - tũliĩ'äk
т*	tt	morphemic boundary	мōты - mō:ttĩ
с*	ss	morphemic boundary	късак - qəssak
сс*	s	morphemic boundary	кōссоын - qoɣo:lin
ш*	šš	morphemic boundary	кушат - kuššat
к*	qq	morphemic boundary	канықын - qaniqqĩn
к	kk	DUR according to morphological rules	нōтықыт - n'ō:tĩkkĩt
п	pp	PST.NAR / HAB according to morphological rules	ēпынты - ε:ppĩntĩ йпыты - i:ppĩtĩ

\* [Kazakevich, Budyanskaya 2010];

\*\* [Kuznecova et al. 1980];

\*\*\* All the amendments with double hard consonants are also actual for soft consonants (*n'* - *n'n'*, *s'* - *s's'* etc)

Northern Selkup dialects: Evgeniya Smorgunova's transcription

Table 6. Automatic conversion of Smorgunova's transcription into INEL transcription for Northern Selkup dialects

Smorgunova's transcription	INEL transcription
<b>Vowels</b>	
а	a
ä	ä
е	e
и	i
о	o
ö	ö
ө	ɔ (U+0254)
у	u
ү	ü
ь	ə (U+0259)
ы	ɨ (U+0268)
э	ə (U+0259)
<b>Consonants</b>	
к	q
ү	q
б	p
в	w
г	g
д	d
ж	ʒ (U+017E)
з	z
й	j
к	k
қ	q
л	l
л'	l'
м	m
н	n
ң	ŋ (U+014B)
ң	ŋ (U+014B)
п	p
р	r
с	s
т	t
т'	č' (U+010D U+02BC)
ф	f
х	h
х̣	q
ц	c
ч	č' (U+010D U+02BC)
ш	š (U+0161)

ш'	š (U+0161)
щ	š (U+0161)
<b>Diacritics</b>	
˘ (U+0304)	: (U+02D0)
б	' (U+02BC)

Southern and Central Selkup dialects: A. I. Kuzmina's transcription

Table 7. Automatic conversion of Kuzmina's transcription into INEL transcription for Southern and Central Selkup dialects

Kuzmina's transcription	INEL transcription
<b>Vowels<sup>9</sup></b>	
а	a
ä	ä
е	e
и	i
о	o
ö	ö
у	u
ÿ	ÿ
ъ	ə (U+0259)
ы	ɨ (U+0268)
э	ɛ (U+025B)
ë	ə (U+0259)
я	a <after C'> ja <after V> 'a <after C>
<b>Consonants</b>	
к	q
ꞥ	q (except for Ket dialects)
б	b
в	v
г	g
д	d
дж	ǰ (U+01EF)
д'ж'	ǰ' (U+01EF U+02BC)
ж	ʒ (U+0292)
жꞥ	ǰ (U+01EF)
з	z
й	j
к	k
л	l
м	m
н	n
ң	ŋ (U+014B)
п	p
р	r
с	s
т	t
т'	č (U+010D) (except for Ket dialects)
тш	č (U+010D)
ф	f
х	x
хꞥ	q
хꞥ	h
ц	c

<sup>9</sup> No automatic conversion was made for uncommon vowels ë, ю.

ч	č' (U+010D U+02BC)
ш	š (U+0161)
ш'	š (U+0161)
щ	š (U+0161)
<b>Latin symbols left unchanged</b>	
ε	ε (U+025B)
ə	ə (U+0259)
ŷ	ŷ (U+0263)
j	j
l	l
w	w
<b>Diacritics</b>	
˘ (U+0304)	: (U+02D0)
б	' (U+02BC)
<b>Ignored symbols</b>	
° (U+030A)	
◦ (U+0325)	
˜ (U+0303)	
ˆ (U+0351)	
˙ (U+0328)	
ˊ (U+2032)	
ˋ (U+2035)	

### Appendix 3. Morpheme glossing labels (tiers ge, gr)

Table 8. Morpheme glossing labels

Gloss	Value
1DU	first person dual
1PL	first person plural
1SG	first person singular
2DU	second person dual
2PL	second person plural
2SG	second person singular
3DU	third person dual
3PL	third person plural
3SG	third person singular
ABL	ablative case
ABL2	ablative case
ABST	abstract noun derivation
ACC	accusative case
ACTN	action nominal
ACTN2	action nominal
ADES	adessive case
ADJZ	adjectivizer
ADV	adverbial case (in combination with case markers)
ADVZ	adverbializer
ALL	allative case
ATTEN	attenuative derivation
AUD	auditive mood
AUGM	augmentative
AUGM2	augmentative
CAP	captative
CAP.ADJZ	captative adjectivizer
CAR	caritive
CAUS	causative derivation
CO	co-affix
COLL	collective number
COM	comitative
COND	conditional mood
CONJ	conjunctive mood
COR	coordinative case
CVB	converb
CVB2	converb
DEB	debitative derivation
DECAUS	decausative derivation
DEF	definite
DETR	detransitive derivation
DIM	diminutive
DRV	unspecified derivation
DU	dual number
DUR	durative derivation
DYA	dyadic marker
EL	elative case
EL2	elative case
EMPH	emphatic

EP	epenthesis
FRQ	frequentative derivation
FUT	future tense
GEN	genitive case
HAB	habitative derivation
HORT	hortative
ILL	illative case
ILL2	illative case
IMP	imperative mood
INCH	inchoative derivation
INDEF	indefinite clitic/particle
INDEF2	indefinite clitic/particle
INDEF3	indefinite clitic/particle
INDEF4	indefinite clitic/particle
INDEF5	indefinite clitic/particle
INF	infinitive
INFER	inferential mood
INSTR	instrumental case
INSTR2	instrumental case
INSTRN	instrumental noun derivation
INTERJ	interjection with unknown meaning
IPFV	imperfective derivation
IPFV2	imperfective derivation
IPFV3	imperfective derivation
IRREAL	irrealis
ITER	iterative derivation
ITER.NUM	iterative numeral
JUSS	jussive
LOC	locative case
MOM	single action derivation
MULO	multiobjective derivation
MULS	multisubjective derivation
MULT	multiaction derivation
NEG	negative
NEG.EX	negative existence
NEG.IMP	negative imperative
NMLZ	nominalization
NOM	nominative case
O	objective conjugation (in combination with person and number markers)
OBL	oblique case (in combination with possessive markers)
OPT	optative mood
ORD	ordinal numeral
PFV	intensive perfective
PL	plural number
POT.FUT	potential future
PROL	prolative case
PROPR	propriative
PST	past tense
PST.NAR	past narrative tense
PTCP.NEC	necessitative participle
PTCP.PRS	present participle
PTCP.PRS2	present participle

PTCP.PST	past participle
QUEST	interrogative
RES	resultative derivation
RFL	reflexive derivation
RFL.PFV	reflexive perfective
S	subjective conjugation (in combination with person and number markers)
SG	singular number
SNGL	singulative
SUBJ	subjunctive mood
SUP	supine
TEMPN	temporal nominal
TEN.UNIT	ten's unit
TR	transitive derivation
TRL	translative case
US	usitative derivation
VBLZ	verbalizer
VOC	vocative case

## Appendix 4. Morphological category tags (tier mc, ps)

Table 9. Morphological categories tags

Label	Description
adj	Adjective
adv	Adverb
adv:case	Adverbial case ending
clit	Clitic
conj	Conjunction
dem	Demonstrative
emphpro	Emphatic personal/possessive pronoun
interj	Interjection
interrog	Interrogative pro-form
n	Noun
n:case	Nominal case endings
n:ins	Insertion in nouns
n:num	Nominal number ending
n:obl.poss	Nominal suffix of oblique possessive declension
n:poss	Nominal suffix of the possessive declension
n:case.poss	Nominal case suffix of the possessive declension
nprop	Proper Noun
num	Numeral
pers	Personal pronoun
pp	Postposition
prep	Preposition
preverb	Preverb
pro	Pronoun
pro:neg	Negative pronoun marker
ptcl	Particle
ptcp	Participle
quant	Quantifier
qv	Question verb
v	Verb
v:inf	Verbal marker of infinitive form
v:inf.poss	Verbal marker of infinitive form with possessive declension
v:ins	Insertion in verbs
v:mood	Verbal mood marker
v:mood.pn	Verbal mood and person marker
v:pn	Verbal person marker
v:tense	Verbal tense marker
v:tense.mood	Verbal tense and mood marker
%%	Unknown