INEL Evenki Corpus

User documentation

Chris Lasse Däbritz, 07.12.2021

1. Introduction

1.1. Objective of the corpus

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

The INEL Evenki corpus at hand fills a gap in the documentation of the indigenous languages of Northern Eurasia and makes possible further descriptions of the language. Evenki is not completely unknown and undescribed (cf. Vasilevich 1948 & 1958, Konstantinova 1964, Nedjalkov 1997 & 2010, Bulatova & Grenoble 1999), however, the corpus can be a valuable tool for both language-specific and typologically oriented research.

1.2. Evenki language

1.2.1. Description

The Evenki language is spoken by 4,802 people (VPN 2010) in the east of the Russian Federation as well as by a couple of thousand people in China and Mongolia. The latter number highly depends on the account of counting, mainly on the question whether Solon is regarded a variety of Evenki or a separate language (Nedjalkov 1997: xxi–xxii). Here, only Evenki spoken in the Russian Federation is dealt with (cf. sections 1.2.3 and 2.4). The territory, where Evenki is spoken, thus, reaches from the river Yenisei to the Pacific Ocean as well as from the Arctic Ocean to the Mongolian steps. Within the Russian Federation, the most compact areas of Evenki settlements are found along the rivers Stony Tunguska and Lower Tunguska (right tributaries of Yenisei), around Lake Baikal and along the Amur river. Genetically, Evenki belongs to the Northern subgroup of the Tungusic language family, its closest relatives being Even (Lamut), Solon and Negidal (Nedjalkov 1997: xix). In spite of its relatively high number of speakers, Evenki definitely can be regarded as endangered. On the one hand, this is due to its wide distribution over whole Eastern Siberia; on the other hand, Evenki is under heavy influence of Russian, and locally even of Sakha (Yakut) and Buryat (Bulatova & Grenoble 1999: 3).

1.2.2. Language codes

ISO 639-3 code: **evn**

Glottolog code: **even1259**

1.2.3. Dialectal subdivisions

Due to its wide geographical distribution, Evenki exhibits a highly developed dialectal division. In the Russian Federation, three major dialect groups are assumed: Northern, Southern and Eastern dialects. The most important criterion to distinguish the dialect groups is the representation of Proto-Evenki **s*, yielding the reflexes /s/, /š/ and /h/ respectively in various dialects.

[1] https://www.slm.uni-hamburg.de/inel/, last access: 03.11.2021.

[2] http://www.akademienunion.de/en/research/the-academies-programme/, last access: 03.11.2021.

[3] http://www.akademienunion.de/en/, last access: 03.11.2021.

*Table 1: Distribution of the reflexes of Proto-Evenki *s in Evenki dialects*

	Northern		Southern		Eastern	
word-initially	/h/	*hulaki*: ‘fox’	/s/ ~ /š/	*sulaki*: ~ *šulaki*: ‘fox’	/s/	*sulaki*: ‘fox’
word-internally (intervocalically)	/h/	*ahi*: ‘woman’	/s/ ~ /š/	*asi*: ~ *aši*: ‘woman’	/h/	*ahi*: ‘woman’

The Northern dialect group is spoken along the river Lower Tunguska as well as to the north of it. Along the river Lower Tunguska, there are the subdialects Erbogachon (upper reaches) and Ilimpi (middle and lower reaches). Historically, the Ilimpi Evenks were migrating also to the left bank of the river Yenisei, what can be proven by the Evenki language island of Sovetskaya Rechka, a settlement to the west of river Yenisei. Historically, Northern Evenki dialects were spoken by Evenki reindeer herders migrating through the huge territory between the river Lower Tunguska in the south and the rivers Pyasina, Kheta and Khatanga on the Taimyr Peninsula in the north. These varieties were basing most probably on the Ilimpi subdialect, but having undergone noticeable Sakha/Dolgan influence. A remainder of this variety is spoken by some speakers around Xantayskoe Ozero (lake Xantay). A large part of the named population, however, settled down along the river Lower Tunguska and became part of the Ilimpi Evenks. Another part of them took part in the ethnogenesis of the Dolgans on the Taimyr Peninsula during the 19th and early 20th century, shifting from Evenki to Dolgan. Their variety is what is called Taimyr Evenki in the INEL Evenki Corpus. Exemplary diagnostic features of the named subdialects are presented in Table 2.

Table 2: Diagnostic features of Northern Evenki dialects

subdialect	diagnostic feature	representation in other dialects
Erbogachon	assimilated *-nn-* and *-ll-* (e.g. *ha*:*-nni* ~ *ha*:*-nnə* 'know.AOR-2SG' and *ollo* ‘fish’)	consonant clusters *-nd-* and *-ld-* (e.g. *ha*:*-ndi* ~ *ha*:*-ndə* 'know.AOR-2SG' and *oldo* ‘fish’)
Ilimpi	*š* / C_, e.g. *tukša-* ‘to run’	*s*, e.g. *tuksa-* ‘to run’
Kh. Ozero	*-pki*: as form of habitual participle	*-wki*:
Taimyr	person-number ending *-w* generalized for all persons	person-number ending *-w* only for 1SG

The Southern dialect group is spoken along the river Stony Tunguska as well as along the river Sym, a left tributary of Yenisei. The main border within the Southern dialect group runs along the river Yenisei: Varieties spoken to the west of it form the subdialect of Sym, varieties spoken to the east of form the subdialects of Stony Tunguska, Nepa, and Tokma. The most diagnostic feature is again the reflex of Proto-Evenki **s*, yielding /š/ in all positions in Sym (e.g. *šulaki*: ‘fox’, *aši*: ‘woman’), but /s/ in other Southern varieties (e.g. *sulaki*: ‘fox’, *asi*: ‘woman’). The subdialect spoken along the Stony Tunguska forms base of the Evenki literary language, described in most accounts of Evenki grammar (Konstantinova 1964, Nedjalkov 1997 & 2010, Bulatova & Grenoble 1999).

The eastern dialect group, finally, is spoken to the east of Lake Baikal and in the Far East of the Russian Federation, that is, along the river Amur, at the shore of the Pacific Ocean as well as on the island of Sakhalin.

Within the INEL project, mostly those varieties from the Northern and Southern dialect group are dealt with, which had or have contacts with other languages within the scope of the project (Dolgan, Selkup, and to a lesser extent Ket). Consequently, all Northern subdialects are included, except for the variety of Sovetskaya Rechka, since this variety is not represented in the consulted sources. As for the Southern dialect group, mainly the Sym subdialect is included into the corpus. Incidentally, some texts from Stony Tunguska and Nepa were included as well.

To sum up, it can be said that the Evenki language exhibits great dialectal differences in all parts of the language system. This leads to the situation that closer varieties are surely mutually intelligible, which does not necessarily hold true for varieties from different dialect groups and different territories. Within the INEL project, the following varieties are dealt with: Northern dialects > Erbogachon, Ilimpi, Xantayskoe Ozero, Taimyr as well as Southern dialects > Sym (and partly Stony Tunguska, Nepa). In the sections 2.2 and 2.4. it will be described in more detail, which material included into the corpus represents which (sub)dialect group.

1.3. Archiving

The corpus comprises source media files (whenever available) along with the annotated transcripts in *EXMARaLDA*[4] transcript formats and metadata descriptions in *EXMARaLDA* Coma format (see section 2.6.6 for details).

The corpus is archived and published by the Research Data Repository of the Universität Hamburg[5] under open-access conditions with Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0).[6]

1.4. Citation

The corpus is to be cited as follows:

Däbritz, Chris Lasse & Gusev, Valentin. 2021. INEL Evenki Corpus. Version 1.0. Publication date 2021-12-31. https://hdl.handle.net/11022/0000-0007-F43C-3. Archived at Universität Hamburg. In: *The INEL corpora of indigenous Northern Eurasian languages.* https://hdl.handle.net/11022/0000-0007-F45A-1

1.5. Project members

Project summary information

The INEL Evenki corpus has been developed within the long-term INEL project ("Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages"), 2016–2033. For an overview of the INEL project, see Arkhipov & Däbritz (2018). The Evenki subproject spanned three years from January 2019 to December 2021.

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

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

Project leader

Prof. Dr. Beáta Wagner-Nagy

Researchers

Dr. Alexandre Arkhipov (Research coordinator)

Dr. Chris Lasse Däbritz (August 2019 – December 2021)

Dr. Valentin Gusev (January – August 2019)

Developers

Anne Ferger (January 2019 – March 2021)

Daniel Jettka (January 2019 – February 2021)

Timm Lehmberg (Technical coordinator)

Elena Lazarenko (since May 2021)

Aleksandr Riaposov (since April 2021)

Student assistants

Anna Barinskaya (December 2019 – December 2021)

Alena Kulikova (April 2020 – December 2021)

Felicitas Otte (May 2019 – February 2020)

Ozan Özdemir (January – December 2019)

Roman Shtefura (October 2019 – March 2021)

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

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

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

1.6. Acknowledgements

1.6.1. 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.

1.6.2. Organizational Support

The following institutions and persons provided additional organizational support to the project. First of all, the *Taymyr House of National Arts* (TDNT; *Таймырский Дом народного творчества*) in Dudinka made some material from their collection available for the project. The following persons were particularly involved:

Lyubov' Yur'evna Popova, TDNT Director

Tat'yana Viktorovna Ruban, TDNT Vice-Director

Nina Semyonovna Kudryakova, TDNT Head of Department of folklore and ethnography

Tat'yana Vasil'evna Bolina, TDNT Leading Methodologist for Evenki folklore & culture

Tat'yana Vasil'evna Bolina additionally gathered some Evenki material herself during 2018 and 2019 and worked as informant during two weeks of fieldwork in Moscow in August 2019 (see below).

The *Institute of Oriental Manuscripts of the Russian Academy of Sciences* (IVR RAN; *Институт восточных рукописей РАН*) in Saint Petersburg made it possible to purchase the scanned manuscripts from the Rychkov archive (see below). The following persons were particularly involved:

Prof. Dr. Irina Fyodorovna Popova, Director of the IOM RAS

Alla Alekseevna Sizova, Secretary of the Custodian Council, IOM RAS

We would also like to thank Elena Klyachko and Karina Mischenkova for sharing materials collected by L. M. Brodskaya.

2. The corpus

2.1. The language(s) of the corpus

2.1.1. Content

The language of content is mostly Evenki speech, in instances of code-switching also some Russian and Dolgan speech.

2.1.2. Annotations

The main language of annotations is English.

Translations of the original text are provided in English, German and mostly Russian (see tiers **fe**, **fr**, **fg**). For texts from written sources (Vasilevich 1936, Anisimov 1936) original translations into Russian are given (see tier **ltr**) as provided in the publication; the main translations in tier **fr** are often identical but sometimes have been edited. In case of the texts from the Rychkov archived, this principally holds true, too, but here the divergences of original and literal translation are often bigger. For texts transcribed from audio data, the literal translation provided by the native speakers during transcription is given in the tier **ltr** as well.

Morpheme glosses in English, German 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, gg, 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.

2.2. Media

The corpus contains both written and audio data. The material of the corpus stems from four different sources:

1) previously published texts with no audio material available (Vasilevich 1936, Anisimov 1936 and Brodskaya 1967)

2) audio files provided by the Taymyr House of National Arts (TDNT)

3a) audio files recorded from T. V. Bolina during fieldwork session in Moscow

3b) audio files collected and recorded by T. V. Bolina in Taimyr

4) texts from the Rychkov archive with no audio material available

2.3. Selection

The selection of the material to be included into the corpus depended on three parameters: 1) availability, 2) complementarity to the Evenki collection of the project *Minority Languages of Siberia as our Cultural Heritage (Siberian Lang)*[7], and 3) possible language contact to other language/varieties dealt with in the INEL project (especially Dolgan and Selkup). That is why the project focuses on material coming from the Erbogachon, Ilimpi, Xantayskoe Ozero, Taimyr (< Northern) and Sym (< Southern) subdialects, leaving the best described Southern varieties spoken along the river Stony Tunguska largely aside.

In the next section, the choice of material will be explained in more detail when describing the content of the corpus.

2.4. Content

The first package included into the corpus comes from the published works of Vasilevich (1936), Anisimov (1936) and Brodskaya (1967). Regardless of the source, there are unfortunately no corresponding audio files.

- From Vasilevich (1936), texts representing the Ilimpi and Erbogachon (< Northern) subdialects as well as the Sym (< Southern) subdialect were chosen. Incidentally, also one text from the Nepa (< Southern) subdialect was included into the corpus. These texts were collected, transcribed and translated by the author herself in the 1920s and 1930s on fieldtrips as well as in Saint Petersburg from Evenki students of the Herzen Institute. All these texts are folklore texts.
- Texts from Anisimov (1936) represent the Stony Tunguska (< Southern) subdialect. Again, the author himself collected, transcribed and translated the texts. Both folklore and narrative texts do occur.
- Texts from Brodskaya (1967) represent the Xantayskoe Ozero (< Northern) subdialect. Again, the author herself collected, transcribed and translated the texts. Both folklore and narrative texts do occur.

The second package included into the corpus consists of transcripts of contemporary Evenki representing the Xantayskoe Ozero (< Northern) subdialect. The material consists mostly of folklore and narrative texts, moreover, there are a handful of songs and conversations. As for the dialectal provenience, the whole material comes from the Khatayskoe Ozero (< Northern) subdialect. In several texts, quite strong Dolgan interference can be observed. Since this package is based on audio recordings, audio files are available for each transcript in the corpus.

- One part of this package was provided by the Taymyr House of National Arts (TDNT). These texts were recorded in the 2000s in Xantayskoe Ozero and Dudinka. In 2019, they were transcribed and translated by Alexandre Arkhipov and Valentin Gusev on a fieldwork trip to Moscow, being assisted by T. V. Bolina.
- The other part of this package was recorded in 2018 and 2019 in Xantayskoe Ozero and Moscow. In 2018, T. V. Bolina herself recorded few texts from native speakers in Xantayskoe Ozero, but the majority of texts was recorded by Alexandre Arkhipov and Valentin Gusev from T. V. Bolina on a fieldwork trip to Moscow. On the same fieldwork trip, the texts were transcribed and translated by Alexandre Arkhipov and Valentin Gusev, being assisted by T. V. Bolina .

The third package included into the corpus comes from the handwritten archive of the Russian ethnographer Konstantin Mixaylovich Rychkov. Rychkov collected the material in the first two decades of the 20th century on the Taymyr peninsula as well as along the middle and upper reaches of the river Yenisei. The manuscripts are being preserved at the *Institute of Oriental Manuscripts of the Russian Academy of Sciences* (IVR RAN;

[7] Given the existence of this project (http://siberian-lang.srcc.msu.ru/en/textspage, last access: 03.11.2021), no Ilimpi texts from settlements such as Sovetskaya Rechka, Tura or Tutonchany were included, neither Stony Tunguska texts from Poligus, Baykit or Surinda were included into the INEL Evenki Corpus.

Институт восточных рукописей РАН) in Saint Petersburg, from where scanned copies were obtained. The material consists of both folklore and narrative texts, additionally some translations of single sentences as well as songs and riddles are included. All texts appear to be transcribed and translated by K.M. Rychkov himself. As can be expected, there are no corresponding audio files to the texts. The handwritten archive is organized is folders, whereby the texts included into the INEL Evenki Corpus come from the following folders:

- Folder 5: The texts from folder 5 represent the Taimyr (< Northern) subdialect. They were most probably collected in 1908/1909, either on K.M. Rychkov's trip from Dudinka to Khatanga, that is across the Taymyr Peninsula, in March–April 1908, or in subsequent months (April 1908 – February 1909) on the rivers of Volochanka, Avam, Xeta and elsewhere in Taymyr. However, this assumption remains a guess inasmuch metadata are completely lacking from this folder. Only one text in the middle (p. 121) has a date (09.04.1911) and place (river Kemchug) mentioned, most likely date and place of rewriting. Linguistically, these texts differ significantly from texts from other (sub)dialects. This can partly be explained by Dolgan influence (e.g. use of instrumental case instead of prolative case, usage of Turkic-like serial verb constructions), but some patterns rather point to language attrition (generalization of 1SG -*w* for all persons, use of bare verbal stem in complement position).
- Folder 6b: The texts from folder 6b represent the Sym (< Southern) subdialect. It is unknown when exactly the material was collected, since metadata are again lacking completely, except the year 1913 on the folder front page which might indicate the date of collection and/or of re-writing from field notes. The year 1913 is taken into the sigle of the relevant communications (see Section 2.6.4).
- Folder 6v: The texts from folder 6v represent the Ilimpi (< Northern) subdialect. In this folder, there are sporadically some metadata, including the date of recording, the vulgo name of the speaker as well as the place of recording.

2.5. Corpus size

The corpus currently contains 251 transcripts of 41 speakers[8] with 9,765 utterances and 47,778 tokens. 69 transcripts are linked with the respective audio files, which make up a total 3 h 58 min 24 sec of audio material.

2.6. Naming conventions

2.6.1. Name of the corpus

The name of the corpus is INEL Evenki corpus.

2.6.2. Orthography conventions in the corpus

All transcripts in the corpus have a tier **st** (source transcription) and/or a tier **stl** (source transcription Latin). This tier represents the text in its original transcription. In the tier **st**, original Cyrillic transcriptions (material from Brodskaya (1967), TDNT and T.V. Bolina as well as from the Rychkov archive[9]) are displayed, whereas the tier **stl** contains original Latin transcriptions (material from Vasilevich (1936) and Anisimov (1936)) and the mechanical transliteration (material from Brodskaya (1967), TDNT and T.V. Bolina as well as from the Rychkov archive), respectively.

In the tiers **ts** and **tx** a Latin-based phonological transcription is used instead of the Cyrillic script. The transcription is based on principles of both IPA and FUT (Finno-Ugric Transcription). Vowel length is marked by <Vː>, i.e. the sign "Modifier Letter Triangular Colon" after the vowel grapheme. Consonant length is indicated by doubling the consonant grapheme. Palatalization is marked by <C'>, i.e. the consonant grapheme with the sign "Modifier Letter Apostrophe". In the Cyrillic source transcriptions, palatalization is largely marked at the subsequent vowel grapheme. Therefore, in the table below one vowel grapheme in INEL Evenki corresponds to two vowel graphemes in the Cyrillic source transcriptions; the first one is used after non-palatal(ized) consonants, the second one is used after palatal(ized) consonants.

Rychkov uses additional Cyrillic letters <џ>, <һ> and <ң>, the first of which corresponds to the palatalized <d'>. Finally, it should be noted that both <г> and <һ> in Rychkov's transcription can be rendered as INEL Evenki <g>, <ɣ> or <ɦ>, depending on the context. Meanwhile, Rychkov's <x> always corresponds to INEL Evenki <h> (see Arkhipov & Däbritz 2021 for a detailed account of Rychkov's graphic system).

[8] This number is insofar misleading as in the texts from Anisimov (1936) and from the Rychkov archive the number and identity of speakers is often unknown (see sections 2.6.6 and 2.9.3).

[9] Although Rychkov's transcription is based on (pre-reform) Cyrillic alphabet, it also uses some Latin graphemes (e.g. <l>, <w> and <j>); see also 2.10.2.2.

Vasilevich (1936) distinguishes [± ATR] for /i/, /u/ and /a/; additionally, in her transcription <ä> is a front variant of /ə/. Given that the Evenki vowel system is not really settled yet (cf. De Boer 1996 for a detailed account), these distinctions are kept in **ts** and **tx** for the Vasilevich texts. Besides this, Vasilevich (1936) uses superscript <^{r}> for marking affricized [d^r] and [t^r] as well as cedilla for marking palatalness/palatalization, e.g. <ņ> for [n’].

Additional diacritics which occur in Rychkov’s transcription but presumably reflect non-distinctive features, as well as stress marks, are generally not displayed in the INEL Evenki transcription, but kept in the tiers **st** respectively. These are: stress mark as in <a´>, umlaut <ä ö ÿ> (unless representing a prevocalic /j/), low caron <s̬> (rare), and macron <ā> (very rare).

In the corpus the Charis SIL font is used. The following characters are used in the transcriptions:

Table 3: INEL Evenki transcription

INEL Evenki	**Vasilevich 1936**	**Anisimov 1936**	**Cyrillic:** **- TDNT** **- T.V. Bolina** **- Brodskaya 1967**	**Cyrillic:** **- Rychkov**	**IPA**
VOWELS					
i	i	i	ы, и	ы, і	i
ị	ị	-	-	-	i̜
u	u	u	у, ю	у, ӱ	u
ụ	ụ	-	-	-	u̜
o	o	o	о, ё	о, ӧ	o
a	a	a	а, я	а, ä	a
ạ	ạ	-	-	-	a̜
ə	ə	ə	э, е	е	ə
ä	ä	-	-	-	æ
eː	ē	e	э	е	eː
CONSONANTS					
p	p	p	п	п	p
b	в	в	б	б	b
t	t	t	т	т	t
t^r	t^r	-	-	т, (тр)	t^r
d	d	d	д	д	d
d’	ḑ	ḑ	д	ц̡	d^j / ɟ
d^r	d^r	-	-	д, (др)	d^r
k	k	k	к	к	k
g	g	g	г	г, h	g
w	w	w	в	w	w
s	s	s	с	с, s	s
h	h	h	х	х	h
ɦ	-	-	-	г, h	ɦ
ɣ	γ	γ	ғ	г, h	ɣ
č	c	c	ч	ч	ʧ
š	ş	-	ш	ш	ʃ
m	m	m	м	м	m
n	n	n	н	н	n
n’	ņ	ņ	н	н`, н	n^j / ɲ
ŋ	ŋ	ŋ	ӈ	ӊ	ŋ

l	l	l	л	l	l
j	j	j	й	j	j
r	r	r	р	р	r

Capitalization and punctuation

Most of the transcription is written with small letters. Only the first letters of sentences (i.e. after a full stop, question mark, exclamation) and the first letters of proper nouns are written with capital letters. Punctuation follows mostly English punctuation rules. Direct speech is indicated with double inverted commas, e.g. *He said: "The weather is fine today."*.

2.6.3. Folder structure

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

Folders with text transcripts, organized by genre:

- "conv" (conversations)
- "flk" (folklore texts)
- "misc" (miscellaneous; e.g. single sentences, riddles)
- "nar" (narrative texts)
- "song" (songs)

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 formats (*.exb and *.exs) and in ISO/TEI standard "Transcription of Spoken Language"[10]
- sound file in WAV (*.wav) and/or MP3 format (for texts with audio source)
- scanned pages (*.pdf) for texts published in print (Vasilevich 1936, Anisimov 1936, Brodskaya 1967) as well as from the Rychkov archive

Supplementary folders:

- "documentation" (contains user documentation)

Individual files:

- "evenki.coma" (main metadata file)

2.6.4. Transcripts

The names of the transcript files have the structure Speaker_DateOfRecording_Title_Genre, i.e. the same as the respective communication code in the metadata (see 2.6.6.1 for details). The segmented transcript files additionally have a "_s" suffix in the end of their name. 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 audio files have the structure Speaker_DateOfRecording_Title_Genre, i.e. the same as the respective communication code in the metadata (see 2.6.6.1 for details).

2.6.6. Metadata

The main metadata file for the corpus is the *evenki.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), date of recording, communication short title, genre abbreviation. These components are joined by underscore ("_").

[10] http://www.iso.org/iso/cataloguedetail.htm?csnumber=37338 last access: 25.11.2021

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 case of the Sym Evenki material from the Rychkov archive (folder 6v), not the date of recording, but the date of re-writing the manuscript is indicated, since the former is unknown. 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, if available in the given source. Otherwise, the title was created by the compilers of the corpus.

The genre abbreviation can have one of the values *conv* (conversation), *flk* (folklore), *misc* (miscellaneous), *nar* (narrative) and *song* (song).

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

Name: ChAD_20180923_BurbotsEvenks_flk

Speaker: ChAD (Chempogir, Antonina Dmitrievna, see 2.6.6.2)

Date of recording: 23.09.2018

Short title: Burbots [and] Evenks

Genre: folklore (*flk*)

2.6.6.2. Speaker codes

The codes for the speakers are made up of one letter pointing at the last name, one letter pointing at the surname and one letter pointing at the patronymic. E.g. BTV stands for Bolina, Tat`yana Vasil`evna (B = Bolina, T = Tat`yana, V = Vasil`evna). There are two exceptions to this pattern: 1) If the initial phoneme of either part of the name is latinized with a digraph, then two letters may occur (e.g. ChAD for Chempogir, Antonina Dmitrievna). 2) If an abbreviation is already assigned to a different speaker in the INEL project, then the last name of the speaker may be expressed by two letters, e.g. TuMD for Turskaya, Minna Dmitrievna. If the name of a speaker is unknown, the speaker code is NN. In order to distinguish unknown speakers from different sources, we use the speaker codes NNA (unknown speaker in Anisimov (1936)), NNR (unknown speaker in Rychkov, folder 5), NNR2 (unknown speaker in Rychkov, folder 6v), and NNR3 (unknown speaker in Rychkov, folder 6b).

2.6.6.3. Abbreviations

The texts in the corpus were collected by different people, both linguists and non-linguists, and the work in the corpus was done by several people. The abbreviations for all those people as used in the corpus metadata are as follows:

Data collectors and editors

AAF: Anisimov, Arkadiy Fyodorovich

BLM: Brodskaya, Larisa Meyerovna

BTV: Bolina, Tat`yana Vasil`evna

RKM: Rychkov, Konstantin Mixaylovich

VGM: Vasilevich, Glafira Makar`evna

Project members

AAV: Arkhipov, Alexandre

WNB: Wagner-Nagy, Beáta

DCh: Däbritz, Chris Lasse

GVY: Gusev, Valentin

Student assistants

KuA: Kulikova, Alena

OF: Otte, Felicitas

ShR: Shtefura, Roman

Language consultants (transcription and translation)

BTV: Bolina, Tat`yana Vasil`evna

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". Files encoded in the ISO/TEI standard for "Transcription of Spoken Language" (file extensions is ".xml") are intended to be used for enhanced interoperability and export.

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

Audio files are provided in Linear PCM WAVE format (file extension ".wav"), with 16 bit depth and 44 100 Hz (recordings provided by TDNT) or 48 000 Hz sampling frequency (recordings made in 2018–2019 in Taymyr and in Moscow), mono or stereo. It should be noted that the TDNT recordings were obtained as MP3 files (see 2.8.2).

For the previously published folklore texts (Vasilevich 1936, Anisimov 1936, Brodskaya 1967) as well as for the texts from the Rychkov archive, the corresponding pages were scanned and 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

There are two main phases during the workflow of the source files. In the first phase, the workflow differs depending on the source type of the respective transcript. The first phase always ends with the import of the file into *SIL Fieldworks Language Explorer* (FLEx)[11] for glossing.

- Published texts from Vasilevich (1936) and Anisimov (1936): Already published texts were scanned with subsequent OCR (in ABBYY Fine Reader) and saved into a text file, where further processing (conversion from Cyrillic into Latin-based INEL transcription, punctuation clean-up, assignment of tier attributes etc.) was done. Then the texts were saved as plain text and imported into FLEx for glossing.
- Published texts from Brodskaya (1967): Due to poor scan quality, which made subsequent OCR impossible, the texts were manually typed into a text file by a student assistant (KuA). There, further processing (conversion from Cyrillic into Latin-based INEL transcription, punctuation clean-up, assignment of tier attributes etc.) was done. Then the texts were saved as plain text and imported into FLEx for glossing.
- Texts obtained from TDNT: The audio files received from the TDNT were transcribed and translated into Russian in *ELAN*[12] by AAV, GVY and T.V. Bolina during a fieldwork trip to Moscow in 2019. They were further edited in ELAN (conversion from Cyrillic into Latin-based INEL transcription, punctuation clean-up, changes to time-alignment and sentence breaks, assignment of speaker attributes, etc.). After that, the files were saved as flextext XML files and imported into FLEx for glossing (the time-alignment and speaker attributes being imported and preserved in FLEx as well).

[11] https://software.sil.org/fieldworks/, last access: 25.11.2021.

[12] https://archive.mpi.nl/tla/elan, last access: 25.11.2021.

- Texts recorded by T.V. Bolina in 2018/2019: Part of the audio files were transcribed and translated into Russian by T.V. Bolina herself in *SayMore,*[13] which saves natively into ELAN format. The other part was transcribed and translated in ELAN by AAV, GVY and T.V. Bolina. Both parts were further edited in *ELAN* (conversion from Cyrillic into Latin-based INEL transcription, punctuation clean-up, changes to time-alignment and sentence breaks, assignment of speaker attributes, etc.). After that, the files were saved as flextext XML and imported into FLEx for glossing (the time-alignment and speaker attributes being imported and preserved in FLEx as well).
- Texts recorded on fieldwork in 2019: The audio files were transcribed and translated into Russian in ELAN by AAV and GVY with the help of T.V. Bolina during a fieldwork trip to Moscow in 2019. Further processing was also done in ELAN (conversion from Cyrillic into Latin-based INEL transcription, punctuation clean-up, changes to time-alignment and sentence breaks, assignment of speaker attributes, etc.). After that, the files were saved as flextext XML files and imported into FLEx for glossing (the time-alignment and speaker attributes being imported and preserved in FLEx as well).
- Texts from Rychkov archive: The scanned texts from the Rychkov archive were partially transcribed manually, partially recognized automatically (with subsequent manual correction) using the Handwritten Text Recognition (HTR) engine provided by the *Transkribus* program.[14] Several HTR models have been trained successively on different amounts of manually transcribed data up to 521 pages. A more detailed account of the HTR workflow can be found in Arkhipov et al. (2021). After the manual correction the transcript was saved into a text file, where further processing (conversion from Cyrillic into Latin-based INEL transcription, punctuation clean-up, assignment of tier attributes etc.) was done. Then the texts were saved as plain text and imported into FLEx for glossing.

In the second phase, the workflow is the same for all transcripts.

- The tiers imported into FLEx are **ts** (main transcription), **st** and **stl** (original Cyrillic or Latin transcription, if exists), **ltr** (original Russian translation), and **nt** (comments).
- The morphological analysis (interlinear glossing) is done in FLEx. This is when all the morpheme-level tiers are created (**mb, mp, ge, gg, gr, mc**), as well as the part-of-speech tier (**ps**). Also the **BOR** tier is filled directly from the FLEx lexicon.
- As soon as glossing is complete, a text is exported from FLEx as FLEXTEXT XML and converted to EXMARaLDA EXB format. During this conversion, the **ref** tier is created which combines communication code and sentence numbering (see below). There are also some changes to the **tx** tier concerning punctuation and to the morpheme-level tiers concerning the representation of zero morphs (see below).
- After that, all further annotating (and editing) is done in the *EXMARaLDA Partitur-Editor*[15] (see also 2.10).

2.8.2. Media files

The sound files provided by TDNT in MP3 format were eventually converted into Linear PCM WAVE files (44 100 Hz sampling frequency, 16 bit depth).

2.8.3. Metadata

The metadata of the corpus are managed in *EXMARaLDA Corpus Manager* (Coma).[16] Information about the metadata of both speakers and communications was provided either by the sources themselves or by T. V. Bolina. Unfortunately, in the case of the Rychkov archive, metadata are very sparse, and sometimes altogether absent.

[13] https://software.sil.org/saymore/, last access: 25.11.2021.

[14] https://readcoop.eu/transkribus/, last access: 25.11.2021; see Kahle et al. (2017).

[15] http://exmaralda.org/en/partitur-editor-en/, last access: 25.11.2021.

[16] http://exmaralda.org/en/corpus-manager-en/, last access: 25.11.2021

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

2.9.1. Naming conventions and content of the metadata

The general metadata about the whole corpus include the corpus name ("INEL Evenki 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).

2.9.2. Communication metadata

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

Description:

- **0a. Title:** Complete title of the communication.
- **0b. Title (RU):** Complete title of the communication in Russian.
- **0c. Title (EV)**: Complete title of the communication in Evenki, if available.
- **1a. Genre:** Abbreviation of the genre of the communication (conv = conversation, flk = folklore, misc = miscellaneous, nar = narrative, song = song). Note that two persons included not necessarily mean that the communication is a conversation: e.g. there are some communications where one person utters four or five sentences and the other person is talking independently, in those cases we name both speakers but specify the genre as *flk* or *nar*.
- **2a. Recorded by**: Abbreviation of the person by whom the communication was recorded (may be both linguists and non-linguists, see 1.5 and 2.4).
- **2b. Date of recording:** Here the date of recording is given (year only).
- **3a. Dialect:** Here the dialect group (Northern vs. Southern) is specified, where the transcript comes from.
- **3b. Subdialect:** Here the subdialect (Erbogachon, Ilimpi, Xantayskoe Ozero, Taimyr; Sym, Nepa) is specified, where the transcript comes from.
- **4. Speaker(s):** Code(s) of the speaker(s).
- **5a. Transcribed by:** Code of the person who did the transcription.
- **5b. Date of transcribing:** The exact date (if it is known) of the transcribing.
- **7a-c. Translation(s):** Abbreviation of the person who did the translation in question (Russian, English, German).
- **8a. Glossed by:** Abbreviation of the person who did the glossing.
- **8b. Glosses checked:** Abbreviation of the person who checked the glossing.
- **9a-d. Annotation(s):** Abbreviation of the person who did the annotation in question (SeR, SyF, IST, BOR/CS; see 2.10).

Location:

- **Country:** The country where the recording took place; this is always Russia.
- **Region:** The region/administrative unit where the recording took place. We indicate the administrative unit at the time of the recording – consequently, it is e.g. *Turukhanskiy Kray* in case of the Taimyr Evenki transcripts from the Rychkov archive, but *Taymyr Dolgano-Nenets District* in case of the Khantayskoe Ozero transcripts recorded by T. V. Bolina.
- **Settlement (LngLat):** The longitude and latitude of the settlement where the recording took place.
- **Settlement:** The settlement where the recording took place. If no exact settlement is known, also the name of a river, a lake or a mountain ridge can be given.

Languages:

- **Language code:** The language code of the communication (*evn* – Evenki; *rus* – Russian).

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

- **1a. Archive (written):** In case of the Rychkov archive, the folder and the page numbers are indicated, where the given transcript's source text can be found.
- **1b. Number of pages:** Here the number of pages of the latter is given.
- **2. Corresp. sound/written:** If a text from the written archive has a counterpart in sound recordings, the degree of correspondence in transcription is mentioned here (yes/no/partly).
- **3a. Published in:** If the text was published, we give the data of the publication**.** In case of texts from Vasilevich (1936), Anisimov (1936) and Brodskaya (1967), also the text number in the volumes is given.
- **3b. Published in (bibtex):** Here, publication data are given in bibtex format.

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 published communications), 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/her sociolinguistic background. However, due to the great variety of communications and speakers, it is not always possible to give detailed speaker metadata. The following information is given as exactly as possible:

Description of speaker:

- **1a. Family name:** Family name of the speaker (Latin script).
- **1b. Family name (RU):** Family name of the speaker (Cyrillic script).
- **2a. Given name:** Given name of the speaker (Latin script).
- **2b. Given name (RU):** Given name of the speaker (Cyrillic script).
- **3a. Patronymic:** Patronymic of the speaker (Latin script).
- **3b. Patronymic (RU):** Patronymic of the speaker (Cyrillic script).
- **4. Clan:** If known, it is indicated here to which Evenki clan the speaker belongs.
- **5a. Alternate names:** If there are different spellings of names or maiden names etc., they are given here (Latin script).
- **5b. Alternate names (RU):** If there are different spellings of names or maiden names etc., they are given here (Cyrillic script).

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

- **1a. Place of birth:** Place of birth of the speaker (Latin script).
- **1b. Place of birth (RU):** Place of birth of the speaker (Cyrillic script).
- **2. Region:** Region where the speaker was born.
- **3. Country**: Country where the speaker was born; this is always Russia.
- **4. Date of birth:** The speaker's date of birth.
- **5. Date of death:** If the speaker already died, the speaker's date of death.
- **6a. Former residences:** Former residences of the speaker (Latin script).
- **6b. Former residences (RU):** Former residence of the speaker (Cyrillic script).
- **7a. Domicile**: Location where the speaker lived at the time of the recording (Latin script).
- **7b**. **Domicile (RU):** Location where the speaker lived at the time of the recording (Cyrillic script).
- **8a. Other information:** If there is other relevant information on the speaker's biography, it is indicated here.
- **8b. Other information (RU):** Russian translation of (8a).

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

- **1a. Education:** Here information on basic education (i.e. school) of the speaker is given (English).
- **1b. Education (RU):** Here information on basic education (i.e. school) of the speaker is given (Russian).
- **2a. Higher education:** If the speaker has had higher education, it is mentioned here (English).

- **2b. Higher education (RU):** If the speaker has had higher education, it is mentioned here (Russian).
- **3a. Occupation:** Here the profession and/or occupation of the speaker is mentioned (English).
- **3b. Occupation (RU):** Here the profession and/or occupation of the speaker is mentioned (Russian).

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

- **1. Ethnicity:** Ethnicity of the speaker.
- **2a. Ethnicity of mother:** Ethnicity of the speaker's mother.
- **2b. Name of mother:** Name of the speaker's mother.
- **3a. Ethnicity of father:** Ethnicity of the speaker's father.
- **3b. Name of father:** Name of the speaker's father.
- **4a. Ethnicity of husband/wife:** Ethnicity of the speaker's husband/wife.
- **4b. Name of husband/wife:** Name of the speaker's husband/wife.
- **5a. Ethnicity of grandparents:** Ethnicity of the speaker's grandparents.
- **5b. Name of grandparents:** Name of the speaker's grandparents.
- **6a. Family:** Other family members.
- **6b. Family (RU):** Other family members (Russian).

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

- **Informant of:** Here it is mentioned with which linguis(s) the speaker worked.

Languages: Here we give the language codes (*evn* notes Evenki, *rus* Russian, *dlg* Dolgan) for the languages the speaker has command of.

- **L1**
    - **1. First language:** The speaker's first language.
    - **2. Dialect:** Dialect of the speaker's first language.
- **L2**
    - **1. Second language:** The speaker's second language.
    - **2. Dialect:** Dialect of the speaker's second language.
- **…**

2.10. Transcription and annotation

At this point it should be remarked that a lot of 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. See also Arkhipov (2020) for general principles of transcription, annotation and translation.

2.10.1. Tier layout

Every annotation tier has a distinct label (see left column in the table) which is shown in the respective EXB file. In case of multi-speaker transcripts, this label is extended with the speaker code, e.g. *ref-BTV* or *tx-BTV*. The following table shows all occurring tiers and gives a short description of them.

Table 4: Overview of annotation tiers

Tier label	**Tier name**	**Description**	**Unit**	**Optionality**
ref	Reference	Text ID + sentence number	sentence	obligatory
st	Source transcription (Cyrillic)	1) original transcription in Brodskaya (1967) 2) original transcription of T. V. Bolina 3) original transcription in the Rychkov material	sentence	optional
stl	Source transcription (Latin)	1) original transcription in Vasilevich (1936) and Anisimov (1936) 2) automatic transliteration of original transcription in Brodskaya (1967) 3) automatic transliteration of original transcription of T. V. Bolina 4) automatic transliteration of original transcription in the Rychkov material	sentence	optional
ts	Text (sentence)	Main transcription	sentence	obligatory
tx	Text (word)	Main transcription segmented by word for interlinearization	word	obligatory
mb	Morpheme breaks	Morpheme breakdown of words	morph	obligatory
mp	Morphophonemes (underlying)	Underlying (lexical) forms of morphemes	morph	obligatory
ge	Gloss (English)	Morpheme glosses (with lexical glosses in English)	morph	obligatory
gg	Gloss (German)	Morpheme glosses (with lexical glosses in German)	morph	obligatory
gr	Gloss (Russian)	Morpheme glosses (with lexical glosses in Russian)	morph	obligatory
mc	Morphological category	Morphological category/part of speech for each morpheme	morph	obligatory
ps	Part of speech	Part of speech for each word	word	obligatory
SeR	Semantic Role	Semantic (thematic) roles for major NPs	word	optional
SyF	Syntactic function	Syntactic functions for predicates and arguments	word	optional
IST	Information status	Information status for major NPs (given/new/accessible)	word	optional
BOR	Borrowing	Borrowings (source language and type)	word	optional
BOR-phon	Borrowing phonology	Phonological adaptations in borrowings	word	optional
BOR-morph	Borrowing morphology	Morphological adaptations in borrowings	word	optional
CS	Code switching	Code switching and calques (source language and type)	group of words	optional
fe	Free translation (English)	Free translation (English)	sentence	obligatory

Tier label	Tier name	Description	Unit	Optionality
fg	Free translation (German)	Free translation (German)	sentence	obligatory
fr	Free translation (Russian)	Free translation (Russian)	sentence	obligatory
ltr	Literal translation (Russian)	1) Original translation in Vasilevich (1936), Anisimov (1936) and Brodskaya (1967) 2) Original translation of T. V. Bolina 3) Original translation in Rychkov material	sentence	optional
nt	Notes	Notes from corpus developer	sentence	optional

2.10.2. Transcription tiers

2.10.2.1. Main transcription tiers (tx, ts)

The transcription tier (tx) is the most important tier in the transcriptions, as it contains the main transcription segmented into words and is the basis for all further annotations. The transcription tier uses the orthography described in 2.6.2. The transcription tier is derived from the tier ts and is the basis for the morpheme breakdown in the tier mb. The following example shows tx tier in a transcript from the Rychkov archive.

(1)

ref	NNR_191X_BrotherSister_flk.007 (001.007)	
tx	Ajawd'am	aɦiji!
fe[17]	I want a woman!	

The transcription tier (ts) contains a transcription of the utterances which is partly phonological, partly phonetic. Not each and every idiosyncratic instance of variation is marked here, but major deviations from a so-called "standard" forms are marked. E.g. the variation of the lexeme *sulaki*: ~ *hulaki*: ~ *šulaki*: 'fox' is taken into account, but not e.g. the phonetic realization [ɔ] ~ [o] ~ [o̯] of the phoneme /o/. Russian words and code-switches are represented the same way, i.e. not transliterated from Standard Russian orthography, e.g. if the lexeme for 'milk' <молоко> is pronounced with Akanye, i.e. [malako], then it is written also as *malako*. However, phonetic details cannot be covered here, so the differences in vowel reduction in immediately pre-stressed syllables and all other syllables are not taken into account. Consonant palatalization in Russian words and code-switches, if pronounced, is indicated consequently.

(2)

ref	NNR_191X_BrotherSister_flk.007 (001.007)	
ts	Ajawd'am aɦiji!	
tx	Ajawd'am	aɦiji!
fe	I want a woman!	

Uncertainties and special events like laughter or pauses are indicated in the transcription according to the *INEL Corpora General Transcription and Annotation Principles* (Arkhipov 2020: Ch. 4).

2.10.2.2. Source transcription (st)

The source transcription tier (st) contains the original Cyrillic version of the text in question, if available. This is relevant in case of the texts transcribed by T. V. Bolina, texts from the Rychkov archive as well as of texts taken from Brodskaya (1967). As for T. V. Bolina's transcriptions and the texts from Brodskaya (1967), they largely follow the modern literary Evenki orthography. In Rychkov's material, some additional Cyrillic and Latin characters appear as well as diacritics, including palatlization (ˋ) and stress (ˊ) marks. See 2.6.2 above, and see Arkhipov & Däbritz (2021) for a detailed account of Rychkov's transcription.

[17] "fe" stands for 'free English translation' (see 2.10.3.17). It is introduced already here in order to make the examples understandable.

(3)

ref	NNR_191X_BrotherSister_flk.007 (001.007)	
st	A´jawӌӓм агii!	
ts	Ajawd’am aɦiji!	
tx	Ajawd’am	aɦiji!
fe	I want a woman!	

2.10.2.3. Latin source transcription (stl)

The Latin source transcription tier (stl) contains the original Latin transcription of the text in question, if available. This is relevant for the transcripts from Vasilevich (1936) and Anisimov (1936). In texts transcribed by T. V. Bolina, those from Brodskaya (1967) as well as the texts from the Rychkov archive an automatical transliteration of the Cyrillic source transcription is given here, while the main transcription tiers contain further manual corrections. The following example shows the stl tier in a transcript from Vasilevich (1936).

(4)

ref	BaN_1930_Hares_flk.001 (001.001)		
stl	Amīkān əmərən mundʳukātkārtıkī.		
ts	Amịːkaːn əmərən mundʳukaːtkaːrtịkịː.		
tx	Amịːkaːn	əmərən	mundʳukaːtkaːrtịkịː.
fe	A bear came to the hares.		

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*).

(5)

ref	NNR_191X_BrotherSister_flk.007 (001.007)	
st	A´jawӌӓм агii!	
ts	Ajawd’am aɦiji!	
tx	Ajawd’am	aɦiji!
fe	I want a woman!	

If there is a multi-speaker transcript, then the sentences are counted for every speaker separately. Moreover, then the speaker code of the respective speaker is once more mentioned between communication code and sentence number. Two subsequent sentences of different speakers can, thus, have e.g. the following information in the reference tier: *YUK_NN_BTV_20180909_WhatToTell_conv.BTV.002 (005)* and the following reply *YUK_NN_BTV_20180909_WhatToTell_conv.YUK.002 (006)*.

2.10.3.2. 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.

(6)

ref	NNR_191X_BrotherSister_flk.007 (001.007)	
tx	Ajawd’am	aɦiji!
mb	ajaw-d’a-m	aɦi-ji
fe	I want a woman!	

2.10.3.3. Morphophonemes (underlying) (mp)

The underlying morphemes tier (mp) shows the deep structure of the morphemes which were separated from each other in **mb**. Stems are, thus, represented here by their lexical entry in the FLEx lexicon. Two basic rules are important here: First, in case of vowel harmonic suffixes (*ə* ~ *a* ~ *o*), the suffix variant with /ə/ was chosen as lexeme form. Second, in case of dialectal divergences, the Stony Tunguska (< Southern) forms were chosen as lexeme forms, since most dictionaries represent these variants. This is especially relevant in case of the dialect distinguishing feature /s/ ~ /š/ ~ /h/. In the following example from Northern Evenki, the form *huručo*, thus, corresponds to the underlying form *suru-čə:*.

(7)

ref	NNR_191X_BrotherSister_flk.027 (001.027)	
tx	Əki	huručo.
mb	əki	huru-čo
mp	əkiːn	suru-čəː
fe	The sister went off.	

Zero morphs are mostly not yet represented in **mp**. However, there is one instances where a zero morph is indicated in **mp**, too. This is the aorist suffix *-rə* in the 1st and 2nd person singular. This suffix does not have a surface representation, but blocks other tense-aspect suffixes in the given slot. Therefore, we decided to indicate the aorist suffix *-rə* in **mp**. In analogy to true zero morphemes (see 2.10.3.4), this morpheme is given in square brackets preceded by a dot. The following chart illustrates this.

(8)

ref	NNR_191X_BrotherSister_flk.007 (001.007)	
tx	Ajawd'am	afiiji!
mb	ajaw-d'a-m	afi-ji
mp	ajaw-d'ə.[rə]-m	asiː-jə
fe	I want a woman!	

2.10.3.4. Gloss (ge, gg and gr)

The gloss tiers (ge, gg and gr) contain the English, German and Russian glossing of the morphemes in **mb** and **mp**. Stems receive their respective lexical glosses in the three languages, while affixes are glossed identically in latin script and mostly according to the Leipzig Glossing Rules[18]. For the list of abbreviations used and the list of affixes occurring in the corpus, see Appendix 1 and Appendix 2, respectively. Glosses for all morphemes within a word are separated with hyphens. Non-overt morphemes are given in square brackets preceded by a dot (e.g. ".[NOM]").

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

- mood – person/number: IMP.2SG (imperative, 2nd person singular)
- mood – person/number – clusivity: IMP.1PL.IN (imperative, 1st person plural, inclusive)
- non-finite form – specification of the form: PTCP.PST (past participle), CVB.PURP (purposive converb) etc.

Alternative meanings are separated by a slash (e.g. DAT/LOC for dative/locative case). If the gloss of a morpheme is uncertain (e.g. missing in grammars), the gloss is preceded by one percent sign (e.g. %CVB.SEQ for the suffix *-mətəmi* in Rychkov texts). Morphemes with unknown meaning are glossed with two percent signs (%%). Morphemes, which apparently are derivational suffixes, but those function cannot be further determined, are glossed with DRV.

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

(9)

ref	NNR_191X_BrotherSister_flk.007 (001.007)	
tx	Ajawd'am	añiji!
mb	ajaw-d'a-m	añi-ji
mp	ajaw-d'ə.[rə]-m	asiː-jə
ge	%want-IPFV.[AOR]-1SG	woman-ACC.INDF
gg	%wollen-IPFV.[AOR]-1SG	Frau-ACC.INDF
gr	%хотеть-IPFV.[AOR]-1SG	женщина-ACC.INDF
fe	I want a woman!	

(10)

ref	NNR_191X_BrotherSister_flk.048 (001.048)			
tx	[...]	ilimətəmi	tupsad'ačo	bir'adulaji.
mb		il-i-mətəmi	tupsa-d'a-čo	bir'a-dula-ji
mp		il-i-mətəmi	tuksa-d'ə-čəː	bira-ləː-wiː
ge		stand.up-EP-%CVB.SEQ	run-IPFV-PTCP.PST.[NOM]	river-LAT-RFL.SG
gg		aufstehen-EP-%CVB.SEQ	rennen-IPFV-PTCP.PST.[NOM]	Fluss-LAT-RFL.SG
gr		встать-EP-%CVB.SEQ	бежать-IPFV-PTCP.PST.[NOM]	река-LAT-RFL.SG
fe	[The sister stood up], having stood up she ran to the river.			

2.10.3.5. Morphological category (mc)

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

Table 5: Tags for lexical stems

Tag	Comment
adj	adjective
adv	adverb
cardnum	cardinal numeral
clit	clitic
collnum	collective numeral
conj	conjunction
dem	demonstrative pronoun
emphpro	emphatic pronoun
interj	interjection
interrog	interrogative pronoun
locn	locational noun
n	noun
nprop	proper noun
onom	onomatopoeia
pers	personal pronoun
posspro	possessive pronoun
post	postposition
pro	pronoun
prt	particle
ptcp	participle
quant	quantifier
v	verb

Table 6: Tags for inflectional categories

Tag	**Comment**
Inflection of nominals	
n:case	case suffix at nouns (also at adjectives, numerals, participles and pronouns)
n:ep	epenthetic vowel at nouns (also at adjectives, numerals, participles and pronouns)
n:num	number suffix at nouns (also at adjectives, numerals, participles and pronouns)
n:poss	possessive suffix at nouns (also at adjectives, numerals, participles and pronouns)
n:eval	evaluative suffix at nouns (also at adjectives, numerals, participles and pronouns)
n:rfl.poss	reflexive/anaphoric suffix at nouns (also at adjectives, numerals, participles and pronouns)
Inflection of verbs	
v:conv.impers	impersonal converb suffix at verbs
v:conv.pers	personal converb suffix at verbs
v:ep	epenthetic vowel at verbs
v:eval	evaluative suffix at verbs
v:imp.pn	imperative and person-number suffix at verbs
v:inf	infinitive suffix at verbs
v:mood1	mood suffix (set 1) at verbs
v:mood2	mood suffix (set 2) at verbs
v:num	number suffix at verbs (converbs)
v:pn1	person-number suffix (set 1) at verbs
v:pn2	person-number suffix (set 2) at verbs
v:tense1	tense suffix (set 1) at verbs
v:tense2	tense suffix (set 2) at verbs

The following chart shows an example of how morpheme classes are represented:

(11)

ref	NNR_191X_BrotherSister_flk.048 (001.048)			
tx	[...]	ilimətəmi	tupsad’ačo	bir’adulaji.
mb		il-i-mətəmi	tupsa-d’a-čo	bir’a-dula-ji
mp		il-i-mətəmi	tuksa-d’ə-čəː	bira-ləː-wiː
ge		stand.up-EP-%CVB.SEQ	run-IPFV-PTCP.PST.[NOM]	river-LAT-RFL.SG
mc		v-v:(ep)-v:conv.impers	v-v>v-v>ptcp.[n:case]	n-n:case-n:rfl.poss
fe	[The sister stood up], having stood up she ran to the river.			

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 tags used are more or less the same as in the morphological category tier **mc**, moreover, there are the tags *aux* (auxiliary verb) and *cop* (copula). The copulas *bi-* and *o*ː- are used for linking any constituent (mostly subject NPs) with a non-verbal predicate. The verb *bi-* can also be used as auxiliary verb. Moreover, there are the negative auxiliary verbs *ə-* and *ətə*ː-, the latter having future time reference; those are also marked as *aux* in the part of speech tier.

(12)

ref	NNR_191X_NirgushkaEmergen_flk.267 (001.267)			
tx	Hi	bid’andə	mini	ədijit!
mb	hi	bi-d’a-ndə	mi-ni	ədi-ji-t
mp	siː	bi-d’əː-ndi	bi-ŋiː	ədiː-wiː-t
ge	you.SG.[NOM]	be-FUT.IMM-2SG	I-ATTR	husband-RFL.SG-INSTR
mc	pers.[n:case]	v-v:tense2-v:pn1	pers-pers>posspro	n-n:rfl.poss-n:case
ps	pers	cop	posspro	n
fe	You’ll be my husband!			

(13)

ref	YUK_2007_Coal_flk.043 (043)			
tx	Tagda[19]	bəjəl	nulgihis’ol	bis’otin, […].
mb		bəjə-l	nulgi-hi-s’o-l	bi-s’o-tin
mp		bəjə-l	nulgiː-sin-čəː-l	bi-čəː-tin
ge		human-PL.[NOM]	wander-INCEP-PTCP.PST-PL.[NOM]	be-PST-3PL
mc		n-n:(num).[n:case]	v-v>v-v>ptcp-n:(num).[n:case]	v-v:tense2-v:pn2
ps		n	ptcp	aux
fe	Then the people wandered away, [Larka's late mother said].			

(14)

ref	BTV_20190822_StoneLakeMountain_flk.072 (072)	
tx	Əkəl	ŋoːləttə.
mb	ə-kəl	ŋoːlət-tə
mp	ə-kəl	ŋəːlət-rə
ge	NEG-IMP2.SG	be.afraid-PTCP.NFUT
mc	v-v:imp:pn	v-v>ptcp
ps	aux	ptcp
fe	Don’t be afraid!	

2.10.3.7. Semantic roles (SeR)

The Semantic Roles tier (SeR) contains the annotation of semantic roles (a.k.a. thematic roles, theta-roles). The annotation is based on GRAID principles (cf. Haig & Schnell 2014), and the annotation scheme used was developed by Beáta Wagner-Nagy and Sándor Szeverényi (Wagner-Nagy et al. 2018: 21ff.) who also made it available for the project. The annotation takes into account form, animacy and semantic role of the referent, the tags are built up according to the scheme <form.animacy:semantic role>. If the referent is expressed by a complex phrase, then the semantic role is tagged at the head of the phrase. Zero referents are tagged per default at the predicate of the sentence. Semantic roles are tagged both in main and in dependent clauses. The following tags for the form of the referent are used:

Table 7: Abbreviations for form of the referent

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

[19] *Tagda* ‘then’ is a Russian code-switch, and therefore neither glossed nor annotated.

pro	pronominal referent
v	verb (non-finite forms in small clauses)

In the category “animacy” human and non-human referents are differentiated. Human referents get the abbreviation <h>, non-human referents get no marking in this category.

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

Table 8: Semantic Roles tagged and their abbreviations

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

The following charts shows some examples of tagging Semantic Roles. In example (15), the agent is covertly realized, thus, tagged at the predicate. The reflexive possessive suffix at *d’uːlaːwị* points to the same referent, therefore it is also tagged as third person referent.

(15)

ref	SaX_1931_Boy_flk.026 (004.009)		
tx	Ta:duk	d’u:la:wị	tuksad’asịnan.
mb	ta:-duk	d’u:-la:-wị	tuksa-d’a-sịn-a-n
mp	tar-duk	d’u:-lə:-wi:	tuksa-d’ə-sin-rə-n
ge	that-ABL	house-LAT-RFL.SG	run-IPFV-INCEP-AOR-3SG
SeR	pro:Time	0.3.h:Poss np:G	0.3.h:A
fe	Then she ran home.		

In example (16), the direct object *hụtəlbə:n* is tagged as theme, since it is effected by the action performed, but does not change its quality or state.

(16)

ref	KI_1931_Charchikan_flk.006 (001.006)		
tx	Nuŋan	hụtəlbə:n	ətəjə:tčəwki.
mb	nuŋa-n	hụtə-l-bə:-n	ətəjə:t-čə-wki
mp	nuŋan-n	hutə-l-wə-n	ətəjə:t-d’ə-wki:
ge	3.[NOM]-3SG	child-PL-ACC-3SG	shepherd-IPFV-PTCP.HAB.[NOM]
SeR	pro.h:A	0.3.h:Poss np.h:Th	
fe	He looked for [the other one’s] children.		

In example (17), finally, the subject *kuŋakar* is tagged as experiencer, since it depends on the emotional verb *aja:w-* ‘to love’, and *nuŋannun* is tagged as comitative, since it refers to another person who accompanies the children when they come. The non-finite verb form *əməmi* is tagged as theme, because it depends on the verb and is effected by the performed action, in this case by the mental state of the subject.

(17)

ref	BTV_20190819_Father_nar.023 (023)			
tx	Kuŋakar	nuŋannun	aja:wuŋkitin	əməmi.
mb	kuŋaka-r	nuŋan-nu-n	aja:w-u-ŋki-tin	əmə-mi
mp	kuŋaka:n-l	nuŋan-nu:n-n	ajaw-i-ŋki-tin	əmə-mi:
ge	child-PL.[NOM]	3-COM-3SG	love-EP-PST.DIST-3PL	come-INF
SeR	np.h:E	pro:Com		v:Th
fe	The children loved coming with together him.			

2.10.3.8. Syntactic function (SyF)

In the Syntactic function tier (SyF) basic syntactic functions (i.e. subject, direct object, predicate) are annotated. The annotation is also based on GRAID principles (Haig & Schnell 2014), and the annotation scheme used was developed by Beáta Wagner-Nagy and Sándor Szeverényi (Wagner-Nagy et al. 2018: 24ff.) who also made it available for the project. Hence, the tags are likewise built up according to the scheme <form.animacy:syntactic function>. Subjects and direct objects are tagged at the head of the respective phrase, zero subjects are tagged at the predicate of the clause. In complex predicates, the auxiliary verb is tagged. Two peculiarities are worth mentioning: First, infinitive forms of verbs can function is subject and object in Evenki; in this case, they are tagged “v:S” and “v:O”, respectively. Second, verbal and participle predicates are separated from each other, since their interplay in Evenki morphosyntax is not fully understood yet, and a separate annotation allows for more concise searches and analyses.

The following tags are used for annotating syntactic functions:

Table 9: Tags for annotating syntactic functions

Abbreviation	Comment
Subject	

pro.h:S	pronominal human subject
pro:S	pronominal non-human subject
np.h:S	nominal human subject
np:S	nominal non-human subject
0.1.h:S	zero/covert first-person human subject
0.2.h:S	zero/covert second-person human subject
0.3.h:S	zero/covert third-person human subject
0.3:S	zero/covert third-person non-human subject
v:S	verbal 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
v:O	verbal object
Predicate	
v:pred	verbal predicate
ptcp:pred	participle predicate
n:pred	nominal predicate
adj:pred	attributive/adjectival predicate
pro:pred	pronominal predicate
ptcl:pred	particle predicate

Moreover, copulas are tagged with the tag *cop*. 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 10: Tags for annotating subordinate clauses

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

The following charts show some examples of tagging syntactic functions.

(18)

ref	BTV_20190819_Father_nar.022 (022)		
tx	Kuŋaka:rwə	əməwusiŋkin	tatkittula.
mb	kuŋa-kaː-r-wə	əmə-wu-si-ŋki-n	tat-kit-tula
mp	kuŋa-kəːn-l-wə	əmə-w-sin-ŋki-n	tati-kit-ləː
ge	child-DIM-PL-ACC	come-TR-FREQ-PST.DIST-3SG	learn-NLOC-LAT
SyF	np.h:O	0.3.h:S v:pred	
fe	He looked for the children in school.		

In example (19), the annotation of a verbal object can be seen.

(19)

ref	BTV_20190819_Father_nar.023 (023)			
tx	Kuŋakar	nuŋannun	ajaːwuŋkitin	əməmi.
mb	kuŋaka-r	nuŋan-nu-n	ajaːw-u-ŋki-tin	əmə-mi
mp	kuŋakaːn-l	nuŋan-nuːn-n	ajaw-i-ŋki-tin	əmə-miː
ge	child-PL.[NOM]	3-COM-3SG	love-EP-PST.DIST-3PL	come-INF
SyF	np.h:S		v:pred	v:O
fe	The children loved coming with together him.			

Example (20) displays the annotation of a subordinate clause, more precisely, a conditional clause.

(20)

ref	YM_1931_Bear_flk.018 (003.003)	
tx	D’iktəwə	əməwrəːkis, […].
mb	d’iktə-wə	əmə-w-rəːk-i-s
mp	d’iktə-wə	əmə-p-raːk-i-s
ge	berry-ACC	come-CAUS-CVB.COND1-EP-2SG
SyF	s:cond	
fe	If you bring berries, [I won't eat anything [other] in future].	

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

Table 11: Parameters for determining information status

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

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

Table 12: Basic tags for annotating information status

Tag	**Comment**
Given referents	
giv-active	given and active referent (i.e. mentioned in the current or last sentence)
giv-inactive	given and inactive referent (i.e. mentioned before the last sentence)
Accessible referents	
accs-sit	referent, accessible through the situation (e.g. having breakfast: “Give me the butter, please.”)
accs-aggr	referent, accessible through the aggregation of other referents (e.g. “*Once 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

Since Evenki is a pro-drop language, many referents are not overtly realized in the sentence. Therefore, the information status of non-overt referents is tagged, too. The tag set remains the same, the prefix <0.> is 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.

Another problem which was dealt with is the issue of direct speech: As it is widely known, direct speech tends to change the perspective of both the hearer and the speaker which has consequences for the discourse status of referents as well. Simply spoken, a referent in direct speech has got an information status within the whole discourse/communication (i.e. for the hearer of the whole communication) and an information status within the micro-discourse made up with the usage of direct speech (i.e. for the hearer of the direct speech). As fine-grade discourse analysis is not the main goal of the project and would be very time-consuming, we decided to tag the information status of referents in direct speech on 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> was proposed by Wagner-Nagy et al. (2018: 29) – according to their guidelines this tag is used when a referent occurs in direct speech (ibid.). Furthermore, so-called utterance predicates are tagged by the tag *quot* and it is distinguished between speech and thought (*quot-sp* vs. *quot-th*) (ibid.). The following examples show how the information status is tagged.

In example (21), the shaman was introduced in the sentences before, and now his outer appearance is described. Here, the shaman is referred to with the personal pronoun *nuŋan*, and tagged as "giv-active". Then, his eyes are described; they were not mentioned yet in the discourse, but are surely accessible to the hearer, since humans usually do have eyes. Therefore *əhalin* is tagged with "accs-inf".

(21)

ref	BTV_20190815_ShamanNyokcho_nar.003 (003)			
tx	Nuŋandu	koŋnomol	əhalin	biso:tin.
mb	nuŋa-n-du	koŋnomo-l	əha-l-i-n	bi-so:-tin
mp	nuŋan-n-du:	koŋnomo-l	ə:sa-l-i-n	bi-čə:-tin
ge	3-3SG-DAT/LOC	black-PL	eye-PL-EP-3SG	be-PST-3PL
IST	giv-active		accs-inf	
fe	He had black eyes.			

In example (22), both the personal pronoun *bi* and the full noun phrase *ge*: *hunat* refer to the same referent. The change of perspective can be shown, since the former item is a first person form, and the latter a third person form. As was described above, this change is indicated by adding "-Q" to the tags of referents within direct speech. Thus, *bi* is tagged with "giv-inactive-Q", and *ge*: *hunat* with "giv-active". The verb, introducing the direct speech, finally is tagged with "quot-sp".

(22)

ref	BTV_20190815_BoyGoldenNape_flk.015 (015)				
tx	"Bi	bakam",	ge:	hunat	gunən.
mb	bi	baka-m	ge:	hunat	gun-ə-n
mp	bi	baka.[rə]-m	ge:	huna:t	gu:n-rə-n
ge	I.[NOM]	find.[AOR]-1SG	second.[NOM]	girl.[NOM]	say-AOR-3SG
IST	giv-inactive-Q			giv-active	quot-sp
fe	He had black eyes.				

2.10.3.10. Borrowing (BOR)

The Borrowing tier (**BOR**) contains the annotation of borrowed lexical items. Both the origin of the item in question and the type of borrowing is annotated. The tags are made up as follows: <LANGUAGE:type>. The annotation is implemented already in the FLEx lexicon and automatically exported to EXMARaLDA. For Evenki there are Russian (RUS), Dolgan (DOLG), Sakha/Yakut (YAK), Nenets (NEN). Since Dolgan and Sakha borrowings can often not be distinguished from each other, the tag DOLG/YAK is used then. For the type of borrowing the following tags are used (cf. also Arkhipov (2020: Ch.5)).

Table 13: Tags for annotating borrowings

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

The following charts show some examples of annotating borrowings and their types:

(23)

ref	NNR_191X_Cossacks_nar.004 (001.004)			
tx	Tar	kaɣakil	huruw'o	hələ,
mb	tar	kaɣak-i-l	huru-w'-o	hələ
mp	tar	kaɣak-i-l	suru-p-čəː	sələ
ge	that.[NOM]	cossack-EP-PL.[NOM]	leave-CAUS-PTCP.PST.[NOM]	iron.[NOM]
BOR		RUS:cult		

tx	n'an	poroh,	n'an	har.
mb	n'an	poroh	n'an	har
mp	n'an	poroh	n'an	saːr
ge	again	powder.[NOM]	again	tobacco.[NOM]
BOR		RUS:cult		NEN:cult
fe	The cossacks brought guns, powder and tobacco.			

(24)

ref	BTV_20190815_TwoShamans_nar.007 (007)			
tx	Minŋi	kərgən	aːsin	oːd'an.
mb	min-ŋi	kərgən	aːsin	oː-d'a-n
mp	bi-ŋiː	kərgən	aːčin	oː-d'əː-n
ge	I-ATTR	family.[NOM]	NEG.EX	become-FUT.IMM-3SG
BOR		DOLG/YAK:core		
fe	My family will disappear.			

2.10.3.11. Borrowing phonology and Borrowing morphology (BOR-Phon & BOR-Morph)

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

Table 4 Annotation panel for phonological processes in borrowings

Tag	Comment
Deletions	
inCdel	initial consonant deletion
inVdel	initial vowel deletion (aphaeresis)
medCdel	medial consonant deletion
medVdel	medial vowel deletion (syncope)
finCdel	final consonant deletion
finVdel	final vowel deletion (apocope)
Insertions	
inVins	initial vowel insertion

medVins	medial vowel insertion
finVins	final vowel insertion
Substitutions	
Csub	consonant substitution
Vsub	vowel substitution
Other	
lenition	lenition (weakening)
fortition	fortition (strengthening)

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

Table 15: Tags for annotating morphological processes in borrowings

Tag	**Comment**
Adaptation strategies	
dir:	direct insertion (i.e. insertion without morphological adaptation)
indir:	indirect insertion (i.e. insertion with morphological adaptation)
parad:	paradigm insertion (i.e. a paradigm borrowed)
Further inflection (in the matrix language)	
:bare	no inflection
:infl	further inflection

The following chart shows some examples of annotating both borrowing phonology and borrowing morphology:

(25)

ref	NNR_191X_Cossacks_nar.004 (001.004)			
tx	Tar	kaɣakil	huruw’o	hələ,
mb	tar	kaɣak-i-l	huru-w’-o	hələ
mp	tar	kaɣak-i-l	suru-p-čəː	sələ
ge	that.[NOM]	cossack-EP-PL.[NOM]	leave-CAUS-PTCP.PST.[NOM]	iron.[NOM]
BOR		RUS:cult		
BOR-Phon		CSub		
BOR-Morph		dir:infl		
tx	n’an	poroh,	n’an	har.
mb	n’an	poroh	n’an	har
mp	n’an	poroh	n’an	sa:r
ge	again	powder.[NOM]	again	tobacco.[NOM]
BOR		RUS:cult		NEN:cult
BOR-Phon		CSub		
BOR-Morph		dir:bare		dir:bare
fe	The cossacks brought guns, powder and tobacco.			

Evenki *kaɣak* ‘cossack’ comes from Russian *kazak*, thus, the middle consonant is substituted. The same applies to the final consonant of Evenki *poroh* ‘powder’ (< Russian *porox*). In case of *kaɣak-i-l*, the plural marker *-l* is added, thus, the loanword is further inflected, whence the tag “dir:infl” is used.

2.10.3.12. 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), some instances of Dolgan (DOLG) are also found. The tag set for the type of code-switch is the following:

Table 16: Tags for annotating code-switching

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

The following charts show examples of the annotation of code-switching.

(26)

ref	YUK_2007_BadPeople_nar.002 (002)			
tx	N’ič’evo,	əwədiwə	n’ič’evo	n’et.
mb		əwədi-wə		
mp		əwədi-wə		
ge		Evenki-ACC		
CS	RUS:int.alt		RUS:int.ins	
fe	Nothing, there is no Evenki stuff [anymore].			

(27)

ref	XUK_2007_SongChildren_song.010 (010)		
tx	Ərkoːkun	hɨld’ɨ͡am	[…].
mb	ər-koːkun		
mp	ər-kəːkuːn		
ge	this.[NOM]-AFCT		
CS		DOLG:int.ins	
fe	Now I’ll go, [in holiday I take a wife].		
nt	[DCh]: "hɨld’ɨ͡am" is Dolgan and means 'I will go' (hɨld’-ɨ͡a-m 'go-FUT-1SG').		

2.10.3.13. Free translation (fe, fg, fr)

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 Dolgan original. The translations follow the common guidelines presented in Arkhipov (2020: Ch.3). The following chart shows an example.

(28)

ref	YM_1931_Fox_flk.031 (003.003)		
tx	Bi	haːmị	dụndəwə.
mb	bi	haː-mị	dụndə-wə
mp	bi	saː.[rə]-m	dundə-wə
ge	I.[NOM]	know.[AOR]-1SG	earth-ACC
fe	I know the land.		
fg	Ich kenne das Land.		
fr	Я знаю землю.		

2.10.3.14. Literal Russian translation (ltr)

The Literal Russian translation tier (**ltr**) contains the original Russian translation of the sentence in question. In case of the texts from Vasilevich (1936), Anisimov (1936) and Brodskaya (1967) this means the published translation. In case of the texts made available by the TDNT and transcribed by T. V. Bolina, the transcriber was instructed to provide a literal (sometimes word-to-word) translation, reflecting the underlying Evenki structure.

In the material from the Rychkov manuscripts, finally, contain an original Russian translation in pre-revolutionary Russian orthography. The following chart shows an example of how literal and free translation may differ in a text from the Rychkov archive.

(29)

ref	NNR_191X_Burujdak_flk.022 (001.022)
ts	Omukon dundaɣin Burujdak isšačo tooki ud’an.
fe	At one place Burujdak saw the trace of an elk.
fr	На одном месте Буруйдак увидел след лося.
ltr	На одномѣ мѣстѣ Бурујда´к увидалъ сохатаго слѣдъ.

2.10.3.15. Notes (nt)

The Notes tier (**nt**) eventually contains notes which clarify the content of the sentence or point at something peculiar in the sentence. The notes begin with the indication of who made the note (abbreviation as listed in 2.6.6.3, in square brackets, followed by a colon). The following chart shows an example of it.

(30)

ref	NNR_191X_Burujdak_flk.007 (001.007)			
tx	Nitəkən	tirɣaniwə	n’aɣəd’ačo	Muɣdi.
mb	nitəkən	tirɣani-wə	n’aɣə-d’a-čo	Muɣdi
mp	nitəkən	tirganiː-wə	n’aɣə-d’ə-čəː	Muɣdi
ge	%every.[NOM]	day-ACC	%bow-IPFV-PTCP.PST.[NOM]	Mugdi
fe	Every day he bowed to Mugdi.			
nt	[DCh]: According to RKM (footnote 1), "Mugdi" is a deity of the relevant Evenki clan.			

References

Anisimov, Arkadij F. 1936. *Rodovoe obščestvo ėvenkov (tungusov)*. Trudy po etnografii 1. Leningrad: Izdatel'stvo Instituta Narodov Severa CIK SSSR im. P. G. Smidoviča.

Arkhipov, Alexandre. 2020. *INEL Corpora General Transcription and Annotation Principles*. Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology 5. University of Szeged, Department of Finno-Ugric Studies & Universität Hamburg, Zentrum für Sprachkorpora: Szeged, Hamburg. https://doi.org/10.14232/wpcl.2020.5

Arkhipov, Alexandre & Barinskaya, Anna & Shtefura, Roman. 2021. Using Handwritten Text Recognition on bilingual Evenki-Russian manuscripts of Konstantin Rychkov. *Scripta & e-Scripta* 21. P. 233–244. http://e-scripta.ilit.bas.bg/archives/year-2021/issue-21/using-handwritten-text-recognition-bilingual-evenki-russian-manuscripts-konstantin-rychkov

Arkhipov, Alexandre & Däbritz, Chris Lasse. 2018. Hamburg corpora for indigenous Northern Eurasian languages. *Tomsk Journal of Linguistics and Anthropology* 3 (21), 9–18. Available online at: https://ling.tspu.edu.ru/en/archive.html?year=2018&issue=3&article_id=7130 [Accessed: 19.08.2019].

Arkhipov, Alexandre V. & Däbritz, Chris Lasse. 2021. Reconstructing phonetics behind the graphic system of Evenki texts from the Rychkov archive. *Rhema* 2021(2), 46–64. DOI: 10.31862/2500-2953-2021-2-46-64.

Brodskaya, Larisa M. 1967. *Sintaksičeskie funkcii pričastij v evenkijskom jakyze: na materiale govora evenkov posëlka "Chantajskoe Ozero" Taymyrskogo Nacional'nogo Okruga.* PhD Thesis. Novosibirsk.

Brykina, Maria, Gusev, Valentin, Szeverényi, Sándor and Wagner-Nagy, Beáta. 2018. *Nganasan Spoken Language Corpus (NSLC)*. Archived in Hamburger Zentrum für Sprachkorpora. Version 0.2. Publication date 2018-06-12. Available online at: http://hdl.handle.net/11022/0000-0007-C6F2-8 [Accessed: 19.08.2019].

Bulatova, Nadezhda & Grenoble, Lenore. 1999. *Evenki.* Languages of the world. Materials 141. München: LINCOM Europa.

De Boer, Elisabeth 1996. Present state of the study of Evenki vowel harmony. In: Stary, Giovanni (ed.). *Proceedings of the 38th Permanent International Altaistic Conference (PIAC)*, 121–133. Wiesbaden: Harrassowitz.

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 (2007), 147–187. https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/deliver/index/docId/2036/file/Kapitel6_07.pdf [Accessed: 19.08.2019].

Haig, Geoffrey & Stefan Schnell. 2014. *Annotations using GRAID (Grammatical relations and animacy in discourse)*, Introduction and guidelinesfor annotators, Version 7.0, Available online at https://opus4.kobv.de/opus4-bamberg/frontdoor/index/index/docId/26235 [Accessed: 19.08.2019].

Kahle, P., Colutto, S., Hackl, G. and Mühlberger, G. 2017. Transkribus – a Platform for Transcription, Recognition and Retrieval of Document Images. *IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2017. 19–24.

Konstantinova, Olga A. 1964. *Evenkijskij jazyk: fonetika, morfologija.* Moscow & Leningrad: Nauka.

Nedjalkov, Igor. 1997. *Evenki.* Descriptive Grammars. London [i.a.]: Routledge.

Nedjalkov, Igor. 2010. *Evenki.* Descriptive Grammars. London [i.a.]: Routledge

Vasilevich, Glafira V. 1936. *Sbornik materialov po ėvenkijskomu (tungusskomu) fol'kloru*. Trudy po fol'kloru 1. Leningrad: Izdatel'stvo Instituta Narodov Severa CIK SSSR im. P. G. Smidoviča.

Vasilevich, Glafira V. 1948. *Ocherki dialektov evenkiyskogo (tungusskogo) yazyka.* Leningrad: Uchpedgiz.

Vasilevich, Glafira V. 1958. *Evenkiysko-russkiy slovar. S prilozheniyami i grammaticheskim ocherkom evenkiyskogo yazyka.* [Ewenki-Russian dictionary. With appendices and a grammar sketch of the Ewenki language]. Moscow: Gosudarstvennoe izdatelstvo inostrannykh i nacionalnykh slovarey.

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

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

Appendix 1. Morpheme glossing labels (**ge**, **gg**, **gr**)

Label	**Meaning**
1	first person
2	second person
3	third person
ABL	ablative
ACC	accusative
ACC.INDEF	indefinite accusative
ADD	additive
ADJZ	adjectivizer
ADVZ	adverbializier
AFCT	affection
ALIEN	alienable
ALL	allative
AND	andative
AOR	aorist
ASP	aspect
ATT	attenuative
ATTR	attributive
AUG	augmentative
CAPT	captative
CAUS	causative
CLIT	clitic
COLL	collective
COLL.DAYS	collective suffix used for days
COLL.HUMAN	collective suffix used for humans
COLL.TENTS	collective suffix used for tents
COM	comitative
COMP	comparative
COND	conditional
CONTAINER	derivational suffix forming nouns which describe referents containing other items
CVB	converb
CVB.COND	conditional converb
CVB.INT	intentional converb
CVB.MOD	modal converb
CVB.PURP	purposive converb
CVB.SEQ	sequential converb
CVB.SIM	simultaneous converb
CVB.TERM	terminal converb / converb of bounds
DAT	dative
DECAUS	decausative
DEMN	derivational suffix forming demonyms
DESID	desiderative
DIM	diminutive
DISJ	disjunction
DISTR	distributive
DRV	unspecified derivational suffix
DU	dual
DUR	durative
EMOT	emotive
EMPH	emphasis
EP	epenthetic vowel
EQ.SIZE	derivational suffix (n > adj) expressing ’big as’
EVAL	evaluative
EVID	evidential

EX	exclusive
F	feminine (occurs only in a Russian code-switch)
FOOD	derivational suffix forming verbs which describe events targeted to items of food
FREQ	frequentative
FUT	future
FUT.DIST	distal/remote future
FUT.IMM	immediate future
HAB	habitual
IMP	imperative
IN	inclusive
INCEP	inceptive
INCH	inchoative
INDEF	indefinite
INF	infinitive
INSTR	instrumental
INTJ	interjection
INTNS	intensive
IPFV	imperfective
IRREAL	irreal
ITER	iterative
LAT	lative
LATE	nominal derivational suffix referring to deceased person
LIM	limitative
LOC	locative
MLTP	multiplicative
MOM	momentaneous
NEG	negation
NEG.EX	existential negation
NACT	action noun
NAG	agent noun
NINSTR	instrumental noun
NLOC	nomen loci
NSTATE	state noun
NMLZ	nominalizer
NMLZ.RES	resultative nominalizer
NOM	nominative
NOM.FUT	nominal future
OLD	nominal derivational suffix referring to items not used anymore
ORD	ordinal numeral
PASS	passive
PEJOR	pejorative
PL	plural
PROB	probabilitive
PROL	prolative
PROPR	propriative
PST	past
PST.DIST	distal/remote past
PTCL	particle
PTCP	participle
PTCP.DEB	debitive participle
PTCP.FICT	fictitious participle
PTCP.INT	intentional participle
PTCP.FUT.IMM	immediate future participle
PTCP.HAB	habitual participle
PTCP.NEC	necessitative participle

PTCP.NFUT	non-future participle (used only in connegative)
PTCP.PRF	perfect participle
PTCP.PRS	present participle
PTCP.PST	past participle
PTCP.RES	resultative participle
Q	interrogative clitic/particle
QNT.DISTR	quantitative distributive
REC	reciprocal
REDUPL	reduplication
RES	resultative
RFL	reflexive
SG	singular
SIM	similative
SOC	sociative
SPRL	superlative
TIME	derivational suffix forming nouns/adverbs referring to time points/spans
TR	transitivizer
VBLZ	verbalizer
VEN	venitive
VOC	vocative

Appendix 2. Evenki morphemes in alphabetical order[20]

Marker	Abbreviation	Function
-Ø (zero)	NOM	nominative case
-Ø (zero)	3SG	third person singular (in pn-ending set 2, occurs in conditional and evidential)
-Ø (zero)	3PL	third person plural (in pn-ending set 1)
	%3SG	third person singular (in pn-ending set 1, irregularly used)
-čəː	PTCP.PST	past participle
	PST	past tense marker
-čəːn	NMLZ	nominalizer
-čəːn	DIM	evaluative suffix with diminutive meaning
	PEJOR	evaluative suffix with pejorative meaning
-čəːn	%DU	dual number (not occurring regularly)
-čiː	ATTR	derivational suffix forming adjectives with attributive meaning
-čil	DRV	underspecified derivational suffix (v > v)
-čukəːn	INTNS	evaluative suffix with intensifying meaning
-də	EMPH	emphatic clitic
	ADD	additive clitic
	NEG	negative clitic
	INDEF	indefinite clitic
-dəː	CVB.PURP	purposive converb
-dəː	VBLZ	verbalizer
-dəːwi	IMP.FUT.2SG	imperative future, second person singular
-dələː	CVB.TERM3	terminal converb / converb of bounds
-di	ADJZ	adjectivizer
-din	EQ.SIZE	derivational suffix, forming nominals which describe referents with ‘big as’
-duː	DAT/LOC	dative-locative case
-duk	ABL	ablative case
-d’e	NMLZ	nominalizer
-d’ə	IPFV	imperfective aspect
-d’ə	DRV	underspecified derivational suffix (n > n)
-d’əː	FUT.IMM	immediate future tense
-d’ək	CVB.COND2	conditional converb
-d’əːk	NLOC.PST	nomen loci referring to place where some event happened earlier
-d’əl	FUT.DIST	distal/remote future tense
-d’əli	CVB.TERM2	terminal converb / converb of bounds
-d’əŋəː	PTCP.FUT	future participle
	FUT	future tense
-ga	DAT/LOC	dative-locative case (occurs not regularly, copied from Dolgan)
-gaː	TR	transitivizer
	DRV	underspecified derivational suffix (v > v)
-gdə	LIM	limitative
	NMLZ	nominalizer
-gdi	ADJZ	adjectivizer
-gə	EP	epenthetic / rhythmic element occurring in songs
-gəčin	SIM	evaluative suffix with similative meaning
-gəli	COM2	comitative
-gən	DEMN	derivational suffix forming demonyms
-giː	TR	transitivizer
-giː	ORD	derivational suffix forming ordinal numerals
-giːl	COLL	collective
	PL	irregular plural suffix used for few nouns
-giːt	ABL2	ablative case

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

-gida	%NLOC	derivational suffix forming nomina loci
-gin	IMP.3SG	imperative mood, third person singular
-gu	DRV	underspecified derivational suffix (v > v)
-gu	Q	interrogative clitic/particle
	%EMPH	emphatic clitic/particle
-guli	COM2	comitative
-gut	ADVZ	adverbialiazer (derivational suffix)
-ɣut	NAG	derivational suffix forming agent nouns (copied from Sakha/Dolgan, where the form is -sIt ~ -hIt)
-hu	DRV	underspecified derivational suffix (v > v)
-i	EP	epenthetic vowel
-iŋu	%NMLZ	nominalizer
-ɨm	1SG	possessive suffix, first person singular (occurs not regularly, copied from Dolgan in texts from Xantayskoe Ozero)
-jə	ACC.INDEF	indefinite accusative case
-jə	VOC	vocative
-jə	NMLZ	nominalizer
-jə	PL	irregularly used plural suffix
	COLL	collective
-ji	DRV	underspecified derivational suffix (v > v)
-jo	%NMLZ	nominalizer
-k	NMLZ	nominalizer
-k	NAG	agent noun
-kaː	NMLZ	nominalizer
-kčəː	DIM	diminutive
-kə	EMPH	emphatic clitic
kəːkəːn	EVAL	evaluative suffix
kəːkuːn	AFCT	evaluative suffix expressing affection
-kəːn	DIM	diminutive
	EMPH	evaluative suffix expressing emphasis
	INTNS	evaluative suffix expressing intensification
-kəːnim	CVB.SEQ1	sequential converb
-kəːt	SIM	similative
-kəːt	ATT	attenuative
-kəl	IMP.2SG	imperative mood, second person singular
-kəl	%DRV	underspecified derivational suffix (n > n)
-kəllu	IMP.2PL	imperative mood, second person plural
-kən	%CVB	underspecified converb
-kənə	INTNS	evaluativev suffix expressing intensification
-kəs	NMLZ	nominalizer
-kiː	NMLZ	nominalizer
-kiːt	NMLZ	nominalizer
-kin	NMLZ	nominalizer
-kin	%EVAL	evaluative suffix
-kit	NLOC	derivational suffix forming nomina loci
-kləː	LAT2	lative case
-kliː	ALL/PROL	allative-prolative case
-knən	CVB.TERM1	terminal converb / converb of bounds
-kša	DRV	underspecified derivational suffix (v > v)
-ksə	COLL	collective
	NMLZ.COLL	collective nominalizer
-ksəː	CVB.SEQ2	sequential converb
-ksən	PTCP.FICT	fictitious participle
-ktala	EMOT	emotive (verbal derivational suffix, copied from Sakha/Dolgan)
-ktə	IMP.1SG	imperative mood, first person singular

-ktə	%CVB	underspecified converb
-ktə	ITER	iterative
-ktə	NMLZ	nominalizer
	ADJZ	adjectivizer
	DRV	underspecified derivational suffix (n > n)
-ktən	DRV	underspecified derivational suffix (n > n)
-ktin	IMP.3PL	imperative mood, third person plural
-kun	AUG	augmentative
-kwun	IMP.1PL.EX	imperative mood, first person plural exclusive
-l	PL	plural
	%EMPH	emphasis (contexts where plural occurs unexpectedly, esp. in Rychkov's folder 5)
-l	INCH	inchoative
	VBLZ	verbalizer
-la	PST.F	past tense, feminine (not regular, but code-switch from Russian)
-lan	DRV	underspecified derivational suffix (n > n)
-lar	PL	plural (not regular, copied from Dolgan, occurs only in Rychkov's material, folder 5)
-lbu	ATT	attenuative
-ldə	DRV	underspecified derivational suffix (v > v)
-ldə	COLL.DAYS	collective suffix attached to cardinal numerals, expresses amount of days
-ldəː	VBLZ	verbalizer
-ldi	SOC	sociative (verbal derivational suffix)
-lə	DRV	underspecified derivational suffix (v > v)
-ləː	VBLZ	verbalizer
-ləː	LAT	lative case
-ləːn	NAG	agent noun
-ləːn	PROPR	propriative
-ləhə	DRV	underspecified derivational suffix (n > n)
-lən	%EMPH	emphatic clitic
-lgan	NINSTR	instrumental noun
-lgə	DRV	underspecified derivational suffix (v > v)
-lgə	ADJZ	adjectivizer
-li	%ASP	underspecified aspectual derivational suffix (v > v)
-li	PTCL	cliticized particle
-liː	PROL	prolative
-lin	NACT	action noun
-lkən	PROPR	propriative
-ltə	ADVZ	adverbialzer (derivational suffix)
	%VBLZ	verbalizer
-ltək	PTCP.FUT.IMM	immediate future participle
-ltu	VBLZ	verbalizer
-luː	VBLZ	verbalizer
-m	1SG	first person singular (pn-ending set 1)
-m	VBLZ	verbalizer
-mak	DRV	underspecified derivational suffix (n > n)
-mali	%COMP	comparative
-mčə	COND	conditional mood
-mə	ADJZ	adjectivizer
	DRV	underspecified derivational suffix (n > n)
	%DRV	underspecified derivational suffix (v > v)
-mə	EVAL	evaluative suffix
-məː	VBLZ	verbalizer
-məːč	REC	reciprocal
-məːk	NMLZ	nominalizer
-məːlčə	ATT	attenuative
	MOM	momentaneous

-məːt	EMPH	evaluative suffix expressing emphasis
-məːčin	PTCP.DEB	debitive participle
-məja	%PTCP.NEC	necessitative participle
-mək	CLIT	clitic
-mək	ADJZ	adjectivizer
-məktə	PTCP.RES	resultative participle
-məlčə	%NMLZ	nominalizer
-məmə	INTNS	intensifying derivational suffix
-mətəmi	%CVB.SEQ	sequential converb (occurs only in Rychkov's material folder 5)
-miː	CVB1	underspecified converb
	INF	infinitive
-miː	PEJOR	pejorative evaluative suffix
-miː	VBLZ.CAPT	captative verbalizer
-mịːja	AUG	augmentative
-mkə	ATT	attenuative
-mki	VBLZ	verbalizer
-mkin	NLOC	derivational suffix forming nomina loci
-mmeːn	CVB2	underspecifiec converb
-mna	DRV	underspecified derivational suffix (v > v)
-mnə	NMLZ.RES	resultative nominalizer
-mnək	%ADVZ	adverbializer (derivational suffix)
	CVB.MOD	modal converb
-mniː	NACT	action noun
-mnin	CVB.SEQ3	sequential converb
-mti	FOOD	verbal derivational suffix describing events that are targeted to food
-mu	DISJ	disjunctive clitic
-muː	DESID	desiderative
-n	3SG	third person singular (possessive suffix, pn-ending set 1 & 2)
	%1SG	first person singular (irregularly used)
	%2SG	second person singular (irregularly used)
-n	VBLZ	verbalizer
	DRV	underspecified derivational suffix (v > v)
-n	NACT	action noun
-nä	%ADJZ	adjectivizer
-nčə	ATT	attenuative
	MOM	momentaneous
-ndi	2SG	second person singular (pn-ending set 1)
	%2PL	second person plural (irregularly used)
-nə	AND	andative
	VEN	venitive
-nə	PTCP.PRF	perfect participle
-nə	CVB.SIM1	simultaneous converb
-nə	%NEG.IMP.2SG	negated imperative mood, second person singular
-nəː	PST2	past tense
-nəː	PROB	probabilitive mood
-nəː	INCH	inchoative
-nəːn	COM3	comitative
-ni	NMLZ	nominalizer
-ni	%ACC	accusative case (copied from Dolgan, irregularly used)
-niː	NMLZ	nominalizer
	COLL.HUMAN	collective suffix attached to cardinal numerals, expresses amount of humans
-nik	%ADVZ	adverbializer (derivational suffix, copied from Sakha/Dolgan, irregularly used)
-nil	PL	plural suffix (used with few kinship terms)
-nŋə	NLOC	derivational suffix forming nomina loci
-nu	COLL.TENTS	collective suffix attached to cardinal numerals, expresses amount of tents

-nuːn	COM	comitative
-nun	LIM	limitative clitic
	EMPH	emphatic clitic
-n’i	%DRV	underspecified derivational suffix (v > v)
-ŋɑː	REDUPL	reduplicating clitic
-ŋɑn	EMPH	emphatic clitic
-ŋəː	NMLZ	nominalizer
-ŋəːt	PTCP.INT	intentional participle
	NOM.FUT	nominal future
-ŋəːt	IMP.1PL.IN	imperative mood, first person plural inclusive
-ŋəsiː	CVB.SIM2	simultaneous converb
-ŋəsu	LATE	derivational suffix forming nouns which which refer to deceased people
-ŋi	VBLZ	verbalizer
	DRV	underspecified derivational suffix (v > v)
-ŋi	ALIEN	alienable possession
-ŋị	%NAG	agent noun
-ŋiː	ATTR	derivational suffix forming adjectives with attributive meaning and possessive pronouns
-ŋki	PST.DIST	distal/remote past
-ŋkil	%PTCP	underspecified participle
-ŋnə	HAB	habitual
-p	1PL.IN	first plural inclusive (pn-ending set 1)
-p	CAUS	causative
-p	NINSTR	instrumental noun
-pču	ADJZ	adjectivizer
	%ADVZ	adverbializer (derivational suffix)
-pkɑː	NMLZ	nominalizer
-pkiː	%ALL	allative case
-pti	TIME	derivational suffix pointing to time points/spans
-ptikiː	%ADJZ	adjectivizer
-ptin	NACT	action noun
-ptu	ADJZ	adjectivizer
-ptun	DRV	underspecified derivational suffix (n > n)
-rə	AOR	aorist
	PTCP.NFUT	non-future participle (used only in connegative position)
-rə	MLTP	multiplicative (forming adverbs from cardinal numerals)
-rə	EMPH	emphatic clitic
-rəː	VBLZ	verbalizer
	DRV	underspecified derivational suffix (v > v)
-rəːk	CVB.COND1	conditional converb
-rəːn	OLD	derivational suffix (n > n) pointing no referents not being used anymore
-rən	STATE	state nouns
-rgə	VBLZ	verbalizer
	DECAUS	decausative
-rguː	EVID.HAB	habitual evidential
-riː	PTCP.PRS	present participle
-riːn	ADJZ	adjectivizer
-riktə	LIM	evaluative suffix expressing limitation
-rkə	EVID	evidential
-rmi	VBLZ	verbalizer
-ru	DRV	underspecified derivational suffix (v > v)
-ruk	CONTAINER	derivational suffix forming nouns, which refer to items containing other items
-s	2SG	second person singular (possessive suffix, pn-ending set 2)
-s	2PL	second person plural (pn-ending set 1)
-s	INCEP	inceptive

-s	NMLZ	nominalizer
-sə	COLL	collective
-sə	DRV	underspecified derivational suffix (v > v)
-səl	PL	plural (used with few nouns referring to humans and animals)
-sin	DUR	durative
	FREQ	frequentative
	INCEP	inceptive
	VBLZ	verbalizer
-skiː	LOC.LAT	locative-lative case, attached only to locational nouns
-ssə	DRV	underspecified derivational suffix (v > v)
-sun	2PL	second person plural (possessive suffix, pn-ending set 2)
-t	DUR	durative
	RES	resultative
-t	DECAUS	decausative
-t	INSTR	instrumental case
	ADVZ	adverbializer (derivational suffix)
-t	1PL.IN	first person plural inclusive (possessive suffix, pn-ending set 2)
-t	VBLZ	verbalizer
-tanə	DISTR	evaluative suffix expressing distribution
-təː	VBLZ	verbalizer
-təj	COM	comitative
-təl	DISTR	distributive numeral
-tə	DISTR	distributive (derivational suffix, v > v)
-tikin	QNT.DISTR	quantifying distributive, translated like ‘every’
-til	PL	plural (used with few kinship terms)
-tin	3PL	third person plural (possessive suffix, pn-ending set 2)
-tkəːn	DIM	diminutive
-tkiː	ALL	allative
-tkuː	SPRL	superlative
-tmər	COMP	comparative
-w	1SG	first person singular (possessive suffix, pn-ending set 2)
-w	%2SG, %3SG, %1PL, %2PL, %3PL	used in Rychkov’s material folder 5 for all person-number values
-w	TR	transitivizer
-w	1PL.EX	first person plural exclusive (pn-ending set 1)
-w	VBLZ	verbalizer
-wạ	%CVB	underspecified converb
-waːn	%FREQ	frequentative
-wə	ACC	accusative
-wə	EMPH	emphatic clitic
-wəl	INDEF	indefinite clitic
	EMPH	emphatic clitic
-wər	RFL.PL	1. reflexive possessive suffix, plural possessor 2. same-subject pn-ending, plural subject
	%EMPH.PL	emphatic clitic, used with plural referents
-wət	VBLZ	verbalizer
-wiː	RFL.SG	1. reflexive possessive suffix, singular possessor 2. same-subject pn-ending, singular subject
	%EMPH	emphatic clitic
-wkə	PTCP.NEC	necessitative participle
-wkəːn	CAUS	causative
-wkəːn	NINSTR	instrumental noun
-wkiː	PTCP.HAB	habitual participle
-wu	PASS	passive

-*wun*	1PL.EX	first person plural exclusive (possessive suffix, pn-ending set 2)
-*wunə*	CVB.INT	intentional converb