INEL Dolgan Corpus

User documentation

Chris Lasse Däbritz, 28.11.2022

1. Introduction

1.1. Objective of the corpus

The present corpus of Dolgan 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 Dolgan corpus at hand fills a gap in the documentation of the indigenous languages of Northern Eurasia and makes possible further descriptions of the language. Dolgan was not wholly unknown and undescribed before; however, empirically founded research was often impossible, so that the corpus can be a valuable tool for both language-specific and typologically oriented studies. Additionally, the INEL Dolgan Corpus forms the base for a descriptive grammar of Dolgan (Däbritz 2022).

1.2. Dolgan language

1.2.1. Description

Dolgan is a Turkic language spoken by ca. 1,000 people (VPN 2010). Its speakers live primarily in the Taymyr Dolgano-Nenets District (i.e. mainly on the Taymyr Peninsula), which belongs administratively to the Krasnoyarsk region of the Russian Federation. A small group of speakers of Dolgan is also found in the Anabar District of the Sakha Republic (Yakutia). Together with its closest relative, Sakha (or Yakut), it forms the North Siberian subbranch of the Siberian branch of the Turkic languages (Johanson 2021: 21–24). For a long time, it was considered a dialect of Yakut. Only in 1985, Ubryatova (1985: 3) stated that Dolgan is a separate language having developed from Sakha (Yakut) under the heavy influence of Evenki, a Tungusic language. Due to the predominance of Russian in all official spheres of life, Dolgan must be regarded as a highly endangered language.

1.2.2. Language codes

ISO 639-3 code: **dlg**

Glottolog code: **dolg1241**

1.2.3. Dialectal subdivisions

There are two main dialects of Dolgan: Upper, or South-(West)ern, Dolgan vs Lower, or North-(East)ern Dolgan (cf. Däbritz 2022: Ch. 1.5.4). However, the differences between the dialects are relatively small and rarely compromise mutual comprehensibility.

There seems to be no clear-cut boundary between Upper and Lower Dolgan, but the varieties instead form a dialect continuum. Settlements to the southwest of Xatanga (Ust`-Avam, Volochanka, Katy`ry`k, Xeta, Novaya, Kresty`) are usually classified as representing Upper Dolgan, whereas the settlements to the northeast of Xatanga represent Lower Dolgan (Zhdanixa, Novory`bnoe, Sy`ndassko, Popigaj). Quite a big group of Dolgans also live in

[1] https://www.slm.uni-hamburg.de/inel/, last access: 14.06.2022.
[2] http://www.akademienunion.de/en/research/the-academies-programme/, last access: 14.06.2022.
[3] http://www.akademienunion.de/en/, last access: 14.06.2022.

Dudinka, the administrative centre of the Taymyr Dolgano-Nenets District; this group consists of speakers from the whole area.

Up to the 1930s, Dolgans were also dwelling around today's town of Norilsk; their language was also an Upper variety and formed the base for Ubryatova's (1985) ground-breaking description *Yazyk Norilskix Dolgan* [The language of the Norilsk Dolgans]. Nowadays, this local variety as such is extinct.

As stated above, Dolgan is linguistically close to Sakha (Yakut); as a rule, the more north-eastwards a Dolgan variety is spoken, the more features it shares with Sakha (Yakut). Thereby, the Anabar dialect of Dolgan, mainly spoken in the settlement Yuryung-Xaya, forms a transitory variety; also, Popigaj Dolgan, the north-easternmost variety of Lower Dolgan, is to some extent transitory.

The texts in the corpus stem only from the "core" area of Dolgan, so Anabar Dolgan is not included here.

1.3. Archiving

The corpus comprises source media files (whenever available), the annotated transcripts in *EXMARaLDA*[4] transcript formats and metadata descriptions in the *EXMARaLDA* Coma format (see Section 2.7 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; Kudryakova, Nina; Stapert, Eugénie. 2022. *INEL Dolgan Corpus*. Version 2.0. Publication date 2022-11-30. https://hdl.handle.net/11022/0000-0007-F9A7-4. 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 Dolgan 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 Dolgan subproject spanned three years, from September 2016 to August 2019, and half a year from January 2022 to June 2022.

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 (IFUU, Universität Hamburg)

Researchers

Dr. Alexandre Arkhipov (Research coordinator; IFUU, Universität Hamburg)

Dr. Chris Lasse Däbritz (IFUU, Universität Hamburg)

Dr. Eugénie Stapert (Visiting scholar June 2017 – August 2017 and June 2019 – July 2019; Institut für Skandinavistik, Frisistik und Allgemeine Sprachwissenschaft (ISFAS), Christian-Albrechts-Universität zu Kiel)

Developers

Timm Lehmberg (Technical coordinator)

Elena Lazarenko (since January 2022)

Aleksandr Riaposov (since January 2022)

Daniel Jettka (September 2016 – August 2019)

[4] http://exmaralda.org/en/, last access: 10.06.2022.
[5] https://www.fdr.uni-hamburg.de/communities/inel, last access: 10.06.2022.
[6] https://creativecommons.org/licenses/by-nc-sa/4.0/, last access: 10.06.2022.

Anne Ferger (April 2017 – August 2019)

Niko Partanen (September 2016 – March 2017)

Student assistants

Olesya Degtyareva (October 2016 – December 2017)

Hannes Klitzing (September – December 2016)

Ozan Özdemir (August 2018 – August 2019)

Alena Kulikova (January 2022 – June 2022)

1.6. Acknowledgements

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.

A large part of the audio files in the corpus was made available by the *Taymyr House of National Arts* (TDNT; *Таймырский Дом народного творчества*) in the years 2016 to 2019.[7] Furthermore, Eugénie Stapert allowed the project to include her fieldwork materials in the corpus.

The following institutions and persons provided organizational support for the project during the main project phase in 2016–2019, including a fieldwork trip to Dudinka in July/August 2017.

Lyubov` Yur`evna Popova, TDNT Director

Tat`yana Viktorovna Ruban, TDNT Vice-Director

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

Institute of the World Culture (IWC) at M.V. Lomonosov Moscow State University, and personally:

Acad. Vyacheslav Vsevolodovich Ivanov (1929–2017), IWC Director

The TDNT materials were transcribed and translated by native speakers of Dolgan between 2017 and 2019:

Nina Semyonovna Kudryakova, who also worked as an editor for transcriptions and translations by other consultants

Svetlana Semyonovna Kudryakova

Egor Kudryakov

Adeya Evdokimovna Eske

Aleksandra Tuprina

Illarion Tuprin

During the fieldwork trip in 2017, the following language consultants helped to transcribe, translate and analyze all kinds of texts from the corpus:

Nina Semyonovna Kudryakova

Anna Alekseevna Barbolina

Vera Polikarpovna Bettu

Galina Sidorovna Chuprina

Adeya Evdokimovna Eske

Yuliya Kupchik

Stepanida Il`inichna Kudryakova

Polina Prokop`evna Uodaj

[7] http://www.tdnt.org/, last access: 10.06.2022.

1.7. New in release 2.0

- 20 glossed transcripts (2864 utterances, 19989 tokens) with 03:33:14 hours of corresponding sound were added (cf. Section 2 for details)
    - 1 transcript (17 utterances, 113 tokens) with 00:01:17 hours of corresponding sound from TDNT
    - 18 transcripts (2475 utterances, 17729 tokens) with 03:31:57 hours of corresponding sound from Eugenie Stapert's material
    - 1 transcript (372 utterances, 2147 tokens) from [FD 2000], representing the Norilsk variety
- 37 audio files with 10:00:36 hours without glossed transcripts (cf. Sections 2.3 and 2.4 for details)
    - 28 files with 08:44:49 hours of sound from TDNT
    - 9 files with 01:15:47 hours of sound from Eugenie Stapert's material
- Corrections of grammatical analyses and glossing according to the findings in Däbritz's (2022) grammar, as well as cross-corpora harmonizations
- Additional corpus-wide annotation of Mongolic borrowings
- Additional corpus-wide annotation of existential, locative and possessive predication
- Corrections in further annotations, translations and metadata

2. The corpus

2.1. The language(s) of the corpus

2.1.1. Content

The content language is mostly Dolgan speech, in instances of code-switching, Russian speech and – in folklore texts – a few instances of Evenki speech.

2.1.2. Annotations

The primary language of annotations is English.

Translations of the original text are provided in English, German, and mostly Russian (see tiers **fe**, **fr**, **fg**). As for texts from the written source [FD 2000], 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. As for texts transcribed from the audio files, the literal translation provided by the native speakers during transcription is given in the same tier (**ltr**).

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 metadata language 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 [FD 2000] with no audio material available, 2) audio files made available by the *Taymyr House of National Arts* (TDNT; *Таймырский Дом народного творчества*) and transcribed by local consultants, 3) audio files and transcriptions from various fieldwork sessions of Eugénie Stapert (Kiel) collected in 2008, 2009 and 2010, 4) audio files from an experiment on social cognition done in 2017 by Eugénie Stapert and Chris Lasse Däbritz.

For user convenience, the corpus is available for download in several versions to allow for limited bandwidth and/or storage space: complete version (full-quality WAV files along with their MP3 versions), mp3-only version (MP3 but no WAV files) and no-audio version (transcripts only).

2.3. Selection

The selection of the material to be included depended on its availability and authenticity. At the beginning of the project, only the texts from [FD 2000] were available, so they were the starting point. Later on, transcripts from the TDNT, Eugénie Stapert's collection, and a few own fieldwork recordings were added. As for the TDNT material, the primary criterion for selecting transcriptions was their authenticity: Wherever possible, natural Dolgan speech (e.g. conversations) was chosen, whereas folklore texts, written stories, songs or poems were rather left out. The non-selected material is nevertheless archived and made accessible, consisting of the following items:

- TDNT material

Communication code	**Available files**
AkPG_KuNS_200X_InterviewRadioJournalist_conv	wav
AnSK_KuNS_200504_Novorybnoe_conv	wav
BeAA_KuNS_200X_AgitbrigadeUstAvam_conv	wav
BoNV_KuNS_200X_InterviewTeacher_conv	wav
PoAE_KuNS_200X_DolganLiteracy_conv	wav
ShLY_KuNS_20020724_Interview_conv	wav
UkKN_KuNS_XXXX_Folklore_conv	wav
VeGS_KuNS_2005_CollegeYears_conv	wav
XuUK_KuNS_200X_TaimyrEvenks_conv	wav
ErSV_19XX_WarPartridgesPikes_flk	wav
KoML_1965_OlonkhoHeroSygaday_flk	wav
LaAA_1964_OldManGirlSister_flk	wav
NN3_XXXX_Tale_flk	wav
SaFS_19XX_SufferingLonelyHero_sng	wav
SuAA_2004_SlaveTugutchut_flk	wav
UkET_2002_Mitrofan_flk	wav
AkEE_19XX_PoemBahyrgas_misc	wav
KaD_200X_PoemCloudberry_misc	wav
KuNS_1999_FolkloreFestival_misc	wav
KuNS_2002_FolkloreFestival1_misc	wav
KuNS_2002_FolkloreFestival2_misc	wav
KuNS_20XX_FestivalNationalCostume_misc	wav
KuNS_2004_70thBirthdayESBettu_misc	wav
UkKN_NN4_XXXX_CustomFatherhood_misc	wav
PoNA_KuNS_XXXX_Journalism_nar	wav
PoNA_199205_MyFriend_nar	wav, ELAN
PoNA_19900422_OldManBachada_nar	wav, ELAN
PoNA_20000428_WarRyabov_nar	wav, ELAN

- Eugenie Stapert´s material:

Communication code	**Available files**
AnIM_2009_SleighParts_misc	wav
AnSP_2009_Song_sng	wav

ChAP_2009_Chum_nar	wav
ChAP_2009_Life_nar	wav
ChAP_2009_Tundra_nar	wav
ChVA_2009_Life_nar	wav
ChVA_2009_Life2_nar	wav
KiPP_2009_Song_sng	wav
KiPP_2009_Song2_sng	wav

2.4. Content

The corpus contains texts/transcripts of various genres, broadly classified as conversation, folklore, miscellaneous, narrative, and song; while not being a separate genre, translations are classified apart from the other genres since their language differs in some respects from the original Dolgan texts. Transcripts labelled as "miscellaneous" can be, e.g. poems, but also complex radio transmissions including the speaker's voice, interviews, folklore performances etc.

The whole material from [FD 2000] and some transcriptions provided by the TDNT are folklore texts. Narratives stem from the transcriptions of the TDNT and Eugénie Stapert's fieldwork collection. Conversations originate mainly from the transcripts of the TDNT representing radio interviews; however, some transcripts from Eugénie Stapert's collection and the recording of the experiment on social cognition are also classified as conversations. Most miscellaneous items come from the yet untranscribed part of the TDNT material. Finally, four texts from the material provided by the TDNT are parts of a Bible translation.

2.5. Corpus size

The glossed corpus currently contains 136 transcripts (17 conversations, 51 folklore texts, 1 miscellaneous, 61 narratives, 2 songs, 4 translations) of 74 speakers with 14,193 utterances and 97,625 tokens. 100 transcripts can be linked with the respective audio file, which makes up a total of 14:15:28 hours of audio material.

2.6. Naming conventions

2.6.1. Name of the corpus

The name of the corpus is INEL Dolgan corpus.

2.6.2. Orthography conventions in the corpus

Most transcripts have a tier **st** or **stl** (source transcription (Latin)). This tier represents the text in the Cyrillic or Latin writing system, respectively. In the case of the texts from [FD 2000], this is the original text from the source. In the case of other files, this is the original transcription of the researchers or native language consultants named in 1.6. In the tiers **ts** and **tx**, a Latin-based phonological transcription is used. The transcription is based on principles of both IPA and FUT (Finno-Ugric Transcription). Vowel length is marked by <V:>, i.e. the sign "Modifier Letter Triangular Colon" after the vowel grapheme. Consonant length is indicated by doubling the consonant grapheme. Diphthongs are marked by <V͡V>, i.e. both components of the diphthong combined with the sign "Combining Double Inverted Breve". Palatalization is marked by <C'>, i.e. the consonant grapheme with the sign "Modifier Letter Apostrophe". In the corpus, the Charis SIL font is used. The following characters are used in the transcriptions:

Table 1: INEL Dolgan transcription

INEL transcription		**IPA correspondence**		**Cyrillic orthography**		**Meaning**
a	at	a	at	а	ат	'horse'
e	ebe	ɛ	ɛbɛ	е	эбэ	'river'
o	ogo	ɔ	ɔgɔ	о	ого	'child'
ö	öl	œ	œl	ө	өл	'to die'
ɨ	ɨŋɨ͡ɲa	ɨ	ɨŋɨ͡ɲa	ы	ыңырыа	'bee'
i	ilim	i	ilim	и	илим	'net'
u	uskaːn	u	uskaːn	у	ускаан	'hare'
ü	üs	y	ys	ү	үс	'three'

ʉa	ʉal	ʉa	ʉal	ыа	ыал	‘neighbour’
ɪe	bɪes	ɪɛ	bɪɛs	ие	биэс	‘five’
uo	kuoska	uɔ	kuɔska	уо	куоска	‘cat’
üö	üös	yœ	yœs	үө	үөс	‘stomach’
p	paŋka:	p	paŋka:	п	паңкаа	‘big tea kettle’
b	bar	b	bar	б	бар	‘to go’
t	taba	t	taba	т	таба	‘reindeer’
d	dogor	d	dɔgɔr	д	догор	‘friend’
k	kutujak	k	kutujak	к	кутуйак	‘mouse’
g	gini	g	gini	г	гини	‘he; she; it’
č	če:lke	ʧ	ʧɛ:lkɛ	ч	чээлкэ	‘white’
d’	d’on	ʄ	ʄɔn	дь	дьон	‘people’
s	üs	s	ys	с	үс	‘three’
h	hahɨl	h	hahɨl	h	һаһыл	‘fox’
l	leŋkej	l	lɛŋkɛj	л	лэңкэй	‘snow owl’
r	ürek	r	yrɛk	р	үрэк	‘river’
m	munnu	m	munnu	м	мунну	‘nose’
n	nu:raj	n	nu:raj	н	нуурай	‘to doze off’
n’	n’a:lagaj	ɲ	ɲa:lagaj	нь	ньаалагай	‘midge’
ŋ	üŋkü:le:	ŋ	yŋky:lɛ:	ң / ҥ	үңкүүлээ / үҥкүүлээ	‘to dance’

Capitalization and punctuation

Most of the transcription is written in 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. The punctuation mostly follows English punctuation rules. Direct speech is indicated with double inverted commas, e.g. *He said: “The weather is fine today.”*.

See Arkhipov (2020: Ch. 2.3) for a concise description.

2.6.3. Folder structure

The entire corpus is contained in the folder “DolganCorpus”, which includes the following files and subfolders.

Folders with text transcripts, organized by genre:

- “conv” (conversations)
- “flk” (folklore)
- “misc” (miscellaneous)
- “nar” (narrative)
- “sng” (songs)
- “transl” (texts translated from Russian into Dolgan)

Each genre folder contains one further subfolder per each communication, named identically to the communication name (see Section 2.6.6). Each communication folder contains several files with the same filename identical to the communication name and different extensions according to the file type (see 2.7 for details on file formats):

- annotated transcript in EXMARaLDA, EXB and EXS formats (*.exb, *.exs)
- sound file in WAV (*.wav) (for texts with audio source)

Supplementary folders:

- “documentation” (contains user documentation)

Individual files:

- “dolgan.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 Section 2.6.6 for details). The segmented transcript files also have a "_s" suffix at 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 and video files have the structure Speaker_DateOfRecording_Title_Genre, i.e. the same as the respective communication code in the metadata (see Section 2.6.6 for details).

2.6.6. Metadata

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

2.6.6.1. Names of communications

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

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

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

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

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

Name: MiXS_1967_SoldierInSecondWorldWar_nar

Speaker: MiXS (Mixailov, Xristofor Semyonovich, see 2.6.6.2)

Date of recording: 1967

Short title: SoldierInSecondWorldWar

Genre: nar (narrative)

2.6.6.2. Speaker codes

The codes for the speakers are made up of two letters pointing at the last name, one letter indicating the surname and one letter indicating the patronymic. E.g. MiXS stands for Mixailov, Xristofor Semyonovich (Mi = Mixajlov, X = Xristofor, S = Semyonovich). If an abbreviation is already assigned to a different speaker, then the speaker's last name is expressed by three letters, e.g. ChuAE for A.E. Chuprin.

2.6.7. Additional files

In the case of three of the yet unglossed items from the TDNT collection, an ELAN transcription is available, which is exceptionally added.

2.6.8. Abbreviations

Both linguists and non-linguists collected the texts in the corpus, and several people did the work in the corpus. The abbreviations for all those people as used in the corpus metadata are as follows:

Data collectors and editors

AkAE: Aksyonova, A.E. (radio journalist at the Taymyr radio station)

AkEE: Aksyonova, Evdokiya Egorovna (radio journalist at the Taymyr radio station, Dolgan poetess, developer of the first Dolgan writing system)

AkPG: Aksyonova, Praskov`ya Gavrilovna (radio journalist at the Taymyr radio station)

AlNA, Alekseev, N.A. (Russian ethnographer and historian)

AsKS: Aslamova, Klavdiya Stepanovna (radio journalist at the Taymyr radio station)

EfPE: Efremov, Prokopij Eliseevich (Yakut folklorist and ethnographer)

KuNS: Kudryakova, Nina Semyonovna (radio journalist at the Taymyr radio station; head of the Department of folklore and ethnography of the TDNT)

PoAA: Popov, Andrej Aleksandrovich (Russian ethnographer)

SuAA: Suzdalova, Antonina Alekseevna (Dolgan folklore activist)

UbEI: Ubryatova, Elizaveta Ivanovna (Russian linguist)

UjNN: Ujgurov, N.N. (participant of fieldwork excursions of P.E. Efremov)

VoMS: Voronkin, M.S. (participant of fieldwork excursions of P.E. Efremov)

XaMP: Xarlampiev, Mark Pavlovich (radio journalist at the Taymyr radio station)

ZeA: Zelenkina, A. (radio journalist at the Taymyr radio station)

ZJ: Ziker, John (American ethnographer, working with Dolgans in the 1990s)

Project members and IFUU staff

AAV: Arkhipov, Alexandre

BrM: Brykina, Maria

DCh: Däbritz, Chris Lasse

LE: Lazarenko, Elena

LV: Lopatina, Valeriia

PN: Partanen, Niko

SE: Stapert, Eugénie

Student assistants

DO: Degtyareva, Olesya

KAY: Kulikova, Alyona Y.

KH: Klitzing, Hannes

Language consultants (transcription and translation)

EsAE: Eske, Adeya Evdokimovna

KuE: Kudryakov, Egor

KuNS: Kudryakova, Nina Semyonovna

KuSS: Kudryakova, Svetlana Semyonovna

TuA: Tuprina, Alexandra

TuI: Tuprin, Illarion

UoPP: Uodaj, Polina Prokop`evna

2.7. Technical formats

2.7.1. Transcripts

The annotated transcripts are delivered in the formats of the EXMARaLDA software suite, all 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). A supplementary “segmented transcription” (EXS) is automatically generated from the basic transcription, 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 extension 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") mono, with 44 100 Hz sampling frequency and 16-bit depth. However, it should be noted that it is not their original format in many cases since the TDNT recordings were mainly analogue and were further digitized and stored as MP3 files (see 2.8.2).

MP3 versions of all sound files are also provided as a light-weight option. A version of the corpus with no audio files included is also provided to save bandwidth and storage space (see 2.2).

2.7.4. Other data

No other data types are provided with the corpus.

2.8. Workflow of the source files

2.8.1. Transcripts

The workflow differs depending on the source type of the respective text.

- Texts from the folklore volume [FD 2000] were scanned with subsequent OCR (in Abbyy Fine Reader) and saved as plain text, then converted to Toolbox text format (aka SIL's Standard Format). The resulting Toolbox files were imported into *SIL Fieldworks Language Explorer* (FLEx) [8] for glossing.
- The audio files received from the TDNT were transcribed and translated into Russian by local consultants in *SayMore*[9], which saves natively into ELAN format. They were further edited in *ELAN*[10] (conversion from Cyrillic into Latin-based INEL transcription, punctuation cleanup, changes to time-alignment and sentence breaks, assignment of speaker attributes, etc.). After that, the files were saved as FLEXTEXT files and imported into FLEx for glossing (the time-alignment and speaker attributes being imported and preserved in FLEx as well).
- The audio/video files from Eugénie Stapert's collection were transcribed in ELAN by Eugénie Stapert with the help of local consultants. Some previously glossed texts (in Toolbox) were re-imported into ELAN. After that, all ELAN files were saved as FLEXTEXT files and imported into FLEx for (re-)glossing.
- The audio files of the experiment on social cognition were transcribed in ELAN by Chris Lasse Däbritz with the help of local consultants. After that, it is likewise saved as FLEXTEXT and imported into FLEx for glossing.

The tiers imported into FLEx are **ts** (main transcription), **st** or **stl** (original Cyrillic/Latin transcription, if it exists), **ltr** (original Russian translation) and **fe** (English free translation, for texts from Eugénie Stapert's collection), and **nt** (comments).

The morphological analysis (interlinear glossing) is done for all transcripts in FLEx. This is where all the morpheme-level tiers are created (**mb, mp, ge, gg, gr, mc**) and the part-of-speech tier (**ps**). For all texts except those from [FD 2000], the **BOR** tier is also 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*[11] (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).

[8] https://software.sil.org/fieldworks/, last access: 19.08.2019.
[9] https://software.sil.org/saymore/, last access: 19.08.2019.
[10] https://tla.mpi.nl/tools/tla-tools/elan/, last access: 19.08.2019.
[11] http://exmaralda.org/en/partitur-editor-en/, last access: 19.08.2019.

2.8.3. Metadata

The corpus metadata is managed in *EXMARaLDA Corpus Manager* (Coma)[12].

The metadata of the communications provided by the TDNT was supplied in an MS Word document, converted into an Excel spreadsheet and manually transferred into Coma.

The metadata for materials from Eugénie Stapert's collection was provided in an Excel spreadsheet and transferred manually into Coma.

2.9. Metadata for the corpus

The metadata of the corpus is 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 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 includes the corpus name ("INEL Dolgan 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.
- **1. Genre:** Abbreviation of the genre of the communication (flk = folklore, nar = narrative, conv = conversation, sng = song, transl = translation); note that two persons included does 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 2.6.8).
- **2b. Date of recording:** Here, the date of recording is given (year only).
- **3. Dialect:** If possible, information on the dialect used by the speaker(s) is given here.
- **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.
- **5d. Time-Aligned by:** Abbreviation of the person who aligned the sound to the transcription.
- **6a. Processed by:** Abbreviation of the person who processed (i.e. all technical work before any linguistic analysis; conversions, OCR, sound clearing etc.) the file.
- **6b. Date of processing:** The exact date (if known) of the processing.
- **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-g. Annotation(s):** Abbreviation of the person who did the annotation in question (SeR, SyF, IST, BOR/CS, Top, Foc, ExLocPoss; see 2.10).

Location:

- **Country:** The country where the recording took place; is always Russia.

[12] http://exmaralda.org/en/corpus-manager-en/, last access: 26.10.2017

- **Region:** The region where the recording took place; is either the Taymyr peninsula (until 1930), Taymyr (Dolgano-Nenets) Autonomous Okrug (1930-2007), Taymyr Dolgano-Nenets District (since 2007).
- **Settlement:** The settlement where the recording took place.
- **Settlement (LngLat):** Geographic coordinates (longitude, latitude) of the settlement

Languages:

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

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

- **1a. Archive (sound):** In the case of the TDNT material, the original disc and track numbers of the file are given here.
- **1b. Start-end time:** Here, the start and ending time of the latter is given, if necessary, i.e. if more than one transcript emerges from one file.
- **2. Published in:** If the text was published, we give the data of the publication**.** This is relevant for the texts from [FD 2000]; here also, the text number in the volume is given.
- **2b. 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 about the speaker and, on the other hand, information on their 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 precisely 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. Vulgo (Dolgan name):** Before getting Russian names, Dolgans had their own names and principles of naming persons; if the Dolgan name of a speaker is known, it is given here.
- **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).

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).

Informant of: Here, it is mentioned with whom the speaker worked. However, only linguists doing linguistic fieldwork with them and not radio journalists are named here.

Ethnicity: Here, information about the ethnicity of the respective speaker and their family members are 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).

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).
- **1c. Place of birth (LngLat):** Geographic coordinates (longitude, latitude) of the place of birth
- **2. Region:** Region where the speaker was born; this is mostly the Taymyr peninsula (until 1930), Taymyr (Dolgano-Nenets) Autonomous Okrug (1930-2007), Taymyr Dolgano-Nenets District (since 2007).
- **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 has 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).
- **7c. Domicile (LngLat):** Geographic coordinates (longitude, latitude) of the settlement
- **8a. Other information:** Further information relevant to the speaker's biography.
- **8b. Other information (RU):** Eventual further information relevant to the speaker's biography (Russian).

Languages: Here, we give the language codes (*dlg* notes Dolgan, *evn* Evenki, *nio* Nganasan, *rus* Russian, *sah* Sakha/Yakut) 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 is the respective user guidelines (Wagner-Nagy et al. 2018). This holds especially true for the annotation principles and annotation schemes for the annotation of semantic roles (SeR), syntactic functions (SyF) and information status (IST), as will be shown in the respective sections.

2.10.1. Tier layout

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

Table 2: Overview of annotation tiers

Tier label	Tier name	Description	Unit	Optionality
ref	Reference	Text ID + sentence number	sentence	obligatory
st	Source transcription	1) cyrillic text from [FD 2000] 2) original transcription of the local consultants	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
Top	Topic	Topic-comment-structure	group of words	optional
Foc	Focus	Focus-background-structure	group of words	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
ExLocPoss	Existential, locative and possessive predication	Patterns of existential, locative and possessive predication	group of words	optional
fe	Free translation (English)	Free translation (English)	sentence	obligatory
fg	Free translation (German)	Free translation (German)	sentence	obligatory

Tier label	Tier name	Description	Unit	Optionality
fr	Free translation (Russian)	Free translation (Russian)	sentence	obligatory
ltr	Literal translation (Russian)	1) Original translation in [FD 2000] 2) Literal translation of the local consultants	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 central tier in the transcripts, as it contains the main transcription segmented into words which is the basis for all further annotations, including the morpheme breakdown in the tier mb. The sentence-level transcription tier (ts) has the same content in a single annotation per sentence. The transcription tiers follow the conventions described in 2.6.2.

The transcription used in the corpus is partly phonological and partly phonetic. Not every idiosyncratic instance of variation is marked here, but major deviations from so-called "standard" forms are marked. E.g. the variation of the lexeme for 'head' *men'i*ː ~ *meji*ː is taken into account, but not, e.g. the phonetic realization [ɔ] ~ [o] ~ [o̞] 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 also written 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 considered. Consonant palatalization in Russian words and code-switches, if pronounced, is indicated consequently. See also Arkhipov (2020: Ch. 2.3).

(1)

ts	Ihilletebit l'it'eraturnaj p'er'edač'anɨ.		
tx	Ihilletebit	l'it'eraturnaj	p'er'edač'anɨ.
fe[13]	We broadcast a literary programme.		

Uncertainties and special events like laughter or pauses are indicated in the transcription, according to Arkhipov (2020: Ch. 4).

2.10.2.2. Source transcription (st) / (stl)

The source transcription tiers (st) and (stl), respectively, contain the original Cyrillic and Latin version of the text in question, if available. In the case of the folklore texts from the volume [FD 2000], it is the original text from the book. In the case of the recordings made available by the TDNT, it is the original transcription provided by native speakers. In either case, this means that Cyrillic script is used so that the tier used is **st** (2). Finally, in some of Eugenie Stapert´s texts, the original Latin transcription is preserved, represented in the tier **stl** (3).

(2)

st	Иһиллэтэбит литературнай передачаны.		
ts	Ihilletebit l'it'eraturnaj p'er'edač'anɨ.		
tx	Ihilletebit	l'it'eraturnaj	p'er'edač'anɨ.
fe	We broadcast a literary programme.		

(3)

stl	ol hïrga dieleri χasta:tïbït			
ts	Ol hɨrga d'ɪeleri kasta:tɨbɨt.			
tx	Ol	hɨrga	d'ɪeleri	kasta:tɨbɨt.
fe	We removed the sledge houses.			

[13] "fe" stands for 'free English translation' (see 2.10.3.17).

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 a dot. The sentences are numbered throughout 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*).

(4)

ref	AsKS_19XX_Amulet_nar.001 (001.001)		
st	Иһиллэтэбит литературнай передачаны.		
ts	Ihilletebit l'it'eraturnaj p'er'edač'anɨ.		
tx	Ihilletebit	l'it'eraturnaj	p'er'edač'anɨ.
fe	We broadcast a literary programme.		

If there is a multi-speaker transcript, the sentences are counted for every speaker separately. Moreover, the speaker code of the respective speaker is once more mentioned between the communication code and sentence number. Two subsequent sentences of different speakers can, thus, have, e.g. the following information in the reference tier: *KiPP_KuNS_200211_LifeChildren_conv.KuNS.072 (001.238)* and the following reply *KiPP_KuNS_200211_LifeChildren_conv.KiPP.167 (001.239)*.

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.

(5)

ref	AsKS_19XX_Amulet_nar.001 (001.001)		
tx	Ihilletebit	l'it'eraturnaj	p'er'edač'anɨ.
mb	ihill-e-t-e-bit	l'it'eraturnaj	p'er'edač'a-nɨ
fe	We broadcast a literary programme.		

2.10.3.3. Morphophonemes (underlying) (mp)

The underlying morphemes tier (mp) shows the deep structure of the morphemes separated from each other in **mb**. Stems are, hence, represented here by their lexical entry in the FLEx lexicon. Affixes are represented in their underspecified morphonological structure. The deep forms are written according to turcological tradition (cf. Johanson 2021: Ch. 2.5) and partly adapted to the requirements of Dolgan (mor)phonology; the following chart shows the usage:

Table 3: Representation of deep phonemes

Underspecified phoneme	**Phonological class**	**Possible realizations**
I	high/closed vowels	ɨ, i, u, ü
A	low/open vowels	a, e, o, ö
B	labial consonants	p, b, m
T (suffix-initially) and L	coronal consonants	t, d, n, l
K (suffix-initially) and G	velar consonants	k, g, ŋ
T (suffix-finally)	voiceless stops	p, t, k
K (suffix-finally)	velar stops	k, g
Č[14]	---	č, d', h, s

(6)

ref	AsKS_19XX_Amulet_nar.001 (001.001)		
tx	Ihilletebit	l'it'eraturnaj	p'er'edač'anɨ.

[14] Č appears only in the suffix -ČIt, marking an agent noun.

mb	ihille-t-e-bit	l'it'eraturnaj	p'er'edač'a-nɨ
mp	ihille:-t-A-BIT	l'it'eraturnaj	p'er'edač'a-nI
fe	We broadcast a literary programme.		

Zero morphs are mostly not yet represented in **mp**. However, there are two instances where zero morphs are indicated in **mp**, too. This is, on the one hand, the suffix -tA in the future tense, third-person singular, or future participle plus a possessive suffix, third-person singular, and on the other hand, the causative suffix *-t*. These suffixes do not have a surface representation but cause (mor)phonological changes in stems or other suffixes. Therefore, we decided to indicate them in **mp**. The following chart illustrates this – here, the causative suffix causes fortition of the suffix-initial -B, but does not occur on the surface structure because the consonant cluster **rtp* would be prohibited due to Dolgan phonotactics:

(7)

ref	KiPP_KuNS_200211_LifeChildren_conv.KiPP.100 (001.139)	
tx	[...] olorpotoktoro	bihigini, [...]
mb	olor-potok-toro	bihigi-ni
mp	olor.[t]-BAtAK-LArA	bihigi-nI
fe	[...] they didn't let us sit, [...]	

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 most often according to the Leipzig Glossing Rules[15]. 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. ".[3SG]").

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

- mood – person/number: IMP.2SG (imperative, second-person singular)
- tense – negation: PST2.NEG (past tense 2, negative)
- (negation) – non-finite form – specification of the form: PTCP.PRS (present participle), NEG.CVB.SIM (negative simultaneous converb) etc.

Alternative meanings are separated by a slash (e.g. DAT/LOC and RECP/COLL). Morphemes with unsure meaning are preceded by one per cent sign (%), and morphemes with completely unknown meaning are glossed with two per cent signs (%%).

(8)

ref	AsKS_19XX_Amulet_nar.001 (001.001)		
tx	Ihilletebit	l'it'eraturnaj	p'er'edač'anɨ.
mb	ihille-t-e-bit	l'it'eraturnaj	p'er'edač'a-nɨ
mp	ihille:-t-A-BIT	l'it'eraturnaj	p'er'edač'a-nI
ge	listen-CAUS-PRS-1PL	literary	programme-ACC
gg	zuhören- CAUS-PRS-1PL	literarisch	Sendung-ACC
gr	слушать-CAUS-PRS-1PL	литературный	передача-ACC
fe	We broadcast a literary programme.		

(9)

ref	AsKS_19XX_Amulet_nar.009 (001.009)		
tx	Ogonn'or	töttörü	kan'ɨspat.

[15] https://www.eva.mpg.de/lingua/resources/glossing-rules.php, last access: 16.06.2022.

mb	ogonn’or	töttörü	kan’ɨs-pat
mp	ogonn’or	töttörü	kan’ɨs-pat
ge	old.man.[NOM]	zurück	look.around-NEG.PRS.[3SG]
gg	alter.Mann.[NOM]	back	sich.umsehen-NEG.PRS.[3SG]
gr	старик.[NOM]	назад	осмотреться-NEG.PRS.[3SG]
fe	The old man does not look back.		

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:

Table 4: Tags for lexical stems

Tag	**Comment**
adj	adjective
adv	adverb
cardnum	cardinal numeral
conj	conjunction
dempro	demonstrative pronoun
emphpro	emphatic pronoun
indfpro	indefinite pronoun
interj	interjection
n	noun
ordnum	ordinal numeral
pers	personal pronoun
posspr	possessive pronoun
post	postposition
propr	proper noun
ptcl	particle
quant	quantifier
que	interrogative pronoun
reflpro	reflexive pronoun
v	verb

Table 5: Tags for inflectional categories

Tag	**Comment**
Inflection of nominals	
n:case	case suffix at nouns (also at adjectives and numerals)
n:ins	epenthetic vowel at nouns (also at adjectives and numerals)
n:num	number suffix at nouns (also at adjectives and numerals)
n:poss	possessive suffix at nouns (also at adjectives and numerals)
n:pred.pn	person-number suffix (predicative ending set) at nouns (also at adjectives and numerals)
pro:case	case suffix at pronouns
pro:ins	epenthetic vowel at pronouns
pro:poss	possessive suffix at pronouns
pro:pred.pn	person-number suffix (predicative ending set) at pronouns
Inflection of verbs	
v:case	case suffix at verbs (non-finite forms)
v:cond.pn	person-number suffix (conditional ending set) at verbs
v:cvb	converb suffix at verbs
v:ins	epenthetic vowel at verbs
v:mood	mood suffix at verbs
v:mood.pn	mood and person-number suffix at verbs

v:neg	negation suffix at verbs
v:num	number suffix at verbs (non-finite forms)
v:poss	possessive suffix at verbs (non-finite forms)
v:poss.pn	person-number suffix (possessive ending set) at verbs
v:pred.pn	person-number suffix (predicative ending set) at verbs
v:ptcp	participle suffix at verbs
v:tense	tense suffix at verbs
Inflection of particles[16]	
ptcl:case	case suffix at particles
ptcl:cond.pn	person-number suffix (conditional ending set) at particles
ptcl:ins	epenthetic vowel at particles
ptcl:mood	mood suffix at particles
ptcl:num	number suffix at particles
ptcl:poss	possessive suffix at particles
ptcl:poss.pn	person-number suffix (possessive ending set) at particles
ptcl:pred.pn	person-number suffix (predicative ending set) at particles

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

(10)

ref	AsKS_19XX_Amulet_nar.001 (001.001)		
tx	Ihilletebit	l'it'eraturnaj	p'er'edač'anɨ.
mb	ihille-t-e-bit	l'it'eraturnaj	p'er'edač'a-nɨ
mp	ihille:-t-A-BIT	l'it'eraturnaj	p'er'edač'a-nI
ge	listen-CAUS-PRS-1PL	literary	programme-ACC
mc	v-v>v-v:tense-v:pred.pn	adj	n-n:case
fe	We broadcast a literary programme.		

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 *bu͡ol-* and *e-* are used for linking any constituent (mostly subject NPs) with a non-verbal predicate. The same verbs, as well as *er-*, can also be used as auxiliary verbs. Moreover, there are a number of verbs in Dolgan which form so-called postverbial constructions (a.k.a. aspectual converb constructions, light verb constructions or serial verb constructions); see also Däbritz (2019, 2022: Ch. 6.4.9). These are also marked as *aux* in the part of speech tier.

(11)

ref	AsKS_19XX_Amulet_nar.060 (001.059)		
tx	Karabi:nɨn	hɨrgaga	ötü:le:bit.
mb	karabi:n-ɨ-n	hɨrga-ga	ötü:-le:-bit
mp	karabi:n-tI-n	hɨrga-GA	ötü:-LA:-BIT
ge	carbine-3SG-ACC	sledge-DAT/LOC	string-VBZ-PST2.[3SG]
mc	n-n:poss-n:case	n-n:case	n-n>v-v:tense-v:pred.pn
ps	n	n	v
fe	He tied his carbine up to the sledge.		

(12)

ref	AsKS_19XX_Amulet_nar.031 (001.030)

[16] Particles are listed separately here, as they can take both "nominal" and "verbal" suffixes.

tx	Egeli͡ek	ete.
mb	egel-i͡ek	e-t-e
mp	egel-IAK	e-TI-tA
ge	bring-PTCP.FUT	be-PST1-3SG
mc	v-v:ptcp	v-v:tense-v:poss.pn
ps	v	aux
fe	He would have brought [it].	

(13)

ref	AsKS_19XX_Amulet_nar.065 (001.064)		
tx	Hir	ürde:n	ispit.
mb	hir	ürde:-n	is-pit
mp	hir	ürde:-An	is-BIT
ge	mountain.[NOM]	get.higher-CVB.SEQ	go-PST2.[3SG]
mc	n-n:case	v-v:cvb	v-v:tense-v:pred.pn
ps	n	v	aux
fe	The mountain got higher.		

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.). They also made it available for the project. The annotation considers the form, animacy and semantic role of the referent, and the tags are built up according to the scheme <form.animacy:semantic role>. If a whole phrase expresses the referent, then the semantic role is tagged at the head of the phrase. In postpositional constructions, the cells of the postposition and its complement are merged. Zero referents are tagged per default at the predicate of the sentence. Semantic roles are tagged both in main and dependent clauses. The following tags for the form of the referent are used:

Table 6: Abbreviations for the 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
pro	pronominal referent

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

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

Table 7: Semantic Roles tagged and their abbreviations

Semantic Role	Abbreviation	Comment
Agent	A	- volitional initiator of the action - the participant who 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

Semantic Role	Abbreviation	Comment
		- does not have 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 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 arguments of unaccusative verbs such as *die, fall*
Possessor	Poss	- entity which owns something - both alienable and in-alienable possession - also inanimate referents (e.g. the top of the mountain)
Recipient	R	- (mostly animate) recipient of transfer of something - addressee of verba dicendi
Source	So	- location or entity where a movement starts (i.e. directional location) - original owner in a transfer of something
Stimulus	St	- stimulus for physical perception, i.e. second actant of verbs like *see*, *hear*, *feel*, but NOT of verbs like *look for*, *listen*
Theme	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 show some examples of tagging Semantic Roles:

(14)

ref	AsKS_19XX_Amulet_nar.001 (001.001)		
tx	Ihilletebit	l’it’eraturnaj	p’er’edač’an~~ɨ~~.
mb	ihille-t-e-bit	l’it’eraturnaj	p’er’edač’a-n~~ɨ~~
mp	ihille:-t-A-BIT	l’it’eraturnaj	p’er’edač’a-nI
ge	listen-CAUS-PRS-1PL	literary	programme-ACC
ps	v	adj	n
SeR	0.1.h:A		np:Th
fe	We broadcast a literary programme.		

(15)

ref	AsKS_19XX_Amulet_nar.098 (001.097)					
tx	An~~ɨ~~	gini	küöl	üstün	ünen	iher.
mb	an~~ɨ~~	gini	küöl	üstün	ün-en	ih-er
mp	an~~ɨ~~	gini	küöl	üstün	ün-An	is-Ar
ge	now	3SG.[NOM]	lake.[NOM]	along	crawl-CVB.SEQ	go-PRS.[3SG]

ps	adv	pers	n	post	v	aux
SeR	adv:Time	pro.h:A	pp:Path			
fe	Now he crawls along the lake.					

(16)

ref	AsKS_19XX_Amulet_nar.128 (001.127)		
tx	Ölüökpün	biler	du:.
mb	öl-üök-pü-n	bil-er	du:
mp	öl-IAK-BI-n	bil-Ar	du:
ge	die-PTCP.FUT-1SG-ACC	know-PRS.[3SG]	MOD
ps	v	v	ptcl
SeR	0.1.h:P	0.3:E	
fe	Apparently, he knows that I will die.		

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.). They also made it available for the project. Hence, the tags are likewise built up according to the scheme <form.animacy:semantic role>. Subjects and direct objects are tagged at the head of the respective phrase, and zero subjects are tagged at the predicate of the clause. For complex verbal predicates, the cells of the main verb and the auxiliary are merged. The following tags are used:

Table 8: Tags for annotating syntactic functions

Abbreviation	**Comment**
Subject	
pro.h:S	pronominal human subject
pro:S	pronominal non-human subject
np.h:S	nominal human subject
np:S	nominal non-human subject
0.1.h:S	zero/covert first-person human subject
0.2.h:S	zero/covert second-person human subject
0.3.h:S	zero/covert third-person human subject
0.3:S	zero/covert third-person non-human subject
Direct Object	
pro.h:O	pronominal human direct object
pro:O	pronominal non-human direct object
np.h:O	nominal human direct object
np:O	nominal non-human direct object
Predicate	
v:pred	verbal predicate
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, and the cells belonging to the subordinate clause are merged. The tags are as follows:

Table 9: 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:

(17)

ref	AsKS_19XX_Amulet_nar.001 (001.001)		
tx	Ihilletebit	l'it'eraturnaj	p'er'edač'anɨ.
mb	ihille-t-e-bit	l'it'eraturnaj	p'er'edač'a-nɨ
mp	ihille:-t-A-BIT	l'it'eraturnaj	p'er'edač'a-nI
ge	listen-CAUS-PRS-1PL	literary	programme-ACC
ps	v	adj	n
SyF	0.1.h:S v:pred		np:O
fe	We broadcast a literary programme.		

(18)

ref	AsKS_19XX_Amulet_nar.128 (001.127)		
tx	Ölu͡ökpün	biler	du:.
mb	öl-u͡ök-pü-n	bil-er	du:
mp	öl-IAK-BI-n	bil-Ar	du:
ge	die-PTCP.FUT-1SG-ACC	know-PRS.[3SG]	MOD
ps	v	v	ptcl
SyF	s:comp	0.3:S v:pred	
fe	Apparently, he knows that I will die.		

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 10: Parameters for determining information status

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

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

Table 11: Basic tags for annotating information status

Tag	**Comment**
Given referents	
giv-active	given and active referent (i.e. mentioned in the current or last sentence)
giv-inactive	given and inactive referent (i.e. mentioned before the last sentence)

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

As Dolgan 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 tag <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 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 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 project's main goal 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, 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*, distinguishing between speech and thought (*quot-sp* vs *quot-th*) (ibid.). The following examples show how the information status is tagged:

(19)

ref	AsKS_19XX_Amulet_nar.001 (001.001)		
tx	Ihilletebit	l'it'eraturnaj	p'er'edač'anɨ.
mb	ihille-t-e-bit	l'it'eraturnaj	p'er'edač'a-nɨ
mp	ihille:-t-A-BIT	l'it'eraturnaj	p'er'edač'a-nI
ge	listen-CAUS-PRS-1PL	literary	programme-ACC
ps	v	adj	n
IST	0.accs-sit		new
fe	We broadcast a literary programme.		

Example (19) is the first sentence of a radio programme. Hence it is accessible through the situation that people are broadcasting the programme; therefore, the referent is tagged as *0.accs-sit*. The information in the type of broadcast is, however, new; therefore, the referent is tagged as *new.*

(20)

ref	AsKS_19XX_Amulet_nar.112 (001.111)				
tx	"Hu͡ok,	ünü͡öm,	ünü͡öm",	etiher	ogonn'or.
mb	hu͡ok	ün-ü͡ö-m	ün-ü͡ö-m	et-i-h-er	ogonn'or
mp	hu͡ok	ün-IAK-m	ün-IAK-m	et-I-s-Ar	ogonn'or
ge	no	crawl-FUT-1SG	crawl-FUT-1SG	say-EP-RECP/COLL-PRS.[3SG]	old.man.[NOM]
ps	ptcl	v	v	v	n
IST		0.giv-active-Q	0.giv-active-Q	quot-sp	giv-active
fe	"No, I will crawl, I will crawl", said the old man.				

The context is that a hunter (the old man) got injured, and a polar fox is following him, hoping the old man will die. In the sentence before, the polar fox said “Lie down.”. Hence, the old man is given-active in the discourse.

2.10.3.10. Topic-comment-structure and Focus-background-structure (Top, Foc)

The Topic-comment tier (Top) and Focus-background tier (Foc) contain the annotation of information structure. The tag set and the annotation principles were developed from the Leipzig Model of Information Structure (LM; cf. Junghanns & Zybatow 2009, Däbritz 2021: Ch. 5). The LM operates in the theoretical framework of the Minimalist Program (cf. Chomsky 1995) and was developed to describe the information structure of Slavic languages. However, it is flexible enough to adapt to other languages and language families. The main idea of the LM is that “information structuring is a pragmatically – through the situation of the communication, the context – determined ordering principle through which elements of the sentence get a certain communicative stress.” (Junghanns & Zybatow 2009: 687). Within the LM, there are two information structural levels: On the one hand, the topic-comment-structure and, on the other hand, the focus-background-structure (Junghanns & Zybatow 2009: 688). That means that topic and focus are not complementary in the sentence, both being the salient component on their respective level. *Topic* is understood as an aboutness topic in an Aristotelic sense, i.e. the part of the sentence what the predication is about. In contrast, *focus* is understood as the part of the sentence important for the speaker (ibid.).

Topics are divided into external topics and internal topics, the former standing outside the syntactic structure of the clause (e.g. *That man – he stole my car.*) and the latter standing inside the syntactic structure of the clause (e.g. *My mother worked for the social services.*). Internal topics can be concrete (i.e. having a clearly identifiable referent) or abstract (i.e. situational, so-called frame-setting topics). A special case of a topic is a contrastive topic (e.g. *My mother worked for the social services, but my father worked at TV.*). The tag set developed for topics is the following:

Table 12: Tags for annotating topics

Tag	**Comment**
External topics	
top.ext	external topic
Internal topics	
top.int.concr	concrete internal topic
top.int.concr.contr	concrete contrastive internal topic
top.int.abstr	abstract internal topic

As topical referents can be deleted, covert topics are tagged with <0.> at the predicate of the clause.

Focus is divided into natural focus (a.k.a. informational focus) and special focus. Within natural foci, it is distinguished between wide, intermediate and narrow focus: A wide focus contains the whole clause, an intermediate focus contains the VP of the clause, and a narrow focus contains a single constituent smaller than the VP. Special foci are contrastive foci (e.g. “Since when do you live in Berlin?” – “I live in Dresden now.”) and verum foci (e.g. “Did you buy butter?” – “Yes, I did.”). The tag set developed for foci is the following:

Table 13: Tags for annotating focus

Tag	**Comment**
Natural focus	
foc.wid	wide natural focus
foc.int	intermediate natural focus
foc.nar	narrow natural focus
Special focus	
foc.contr	contrastive focus
foc.ver	verum focus

As only topic and focus are salient features and comment and background can be derived subtractively, only the former are tagged. All the cells belonging to the topic or the focus domain are merged here. The following charts show some examples of annotating information structure:

(21)

ref	AsKS_19XX_Amulet_nar.065 (001.064)		
tx	Hir	ürde:n	ispit.

mb	hir	ürde:-n	is-pit
mp	hir	ürde:-An	is-BIT
ge	mountain.[NOM]	get.higher-CVB.SEQ	go-PST2.[3SG]
Top	top.int.concr		
Foc		foc.int	
fe	The mountain got higher.		

(22)

ref	AsKS_19XX_Amulet_nar.001 (001.001)		
tx	Ihilletebit	l'it'eraturnaj	p'er'edač'anɨ.
mb	ihille-t-e-bit	l'it'eraturnaj	p'er'edač'a-nɨ
mp	ihille:-t-A-BIT	l'it'eraturnaj	p'er'edač'a-nI
ge	listen-CAUS-PRS-1PL	literary	programme-ACC
Top	0.top.int.abstr		
Foc	foc.wid		
fe	We broadcast a literary programme.		

(23)

ref	AsKS_19XX_Amulet_nar.112 (001.111)				
tx	"Huo͡k,	ünüö͡m,	ünüö͡m",	etiher	ogonn'or.
mb	huo͡k	ün-üö͡-m	ün-üö͡-m	et-i-h-er	ogonn'or
mp	huo͡k	ün-IAK-m	ün-IAK-m	et-I-s-Ar	ogonn'or
ge	no	crawl-FUT-1SG	crawl-FUT-1SG	say-EP-RECP/COLL-PRS.[3SG]	old.man.[NOM]
Top		0.top.int.concr	0.top.int.concr		top.int.concr
Foc	foc.ver	foc.contr	foc.contr	foc.int	
fe	"No, I will crawl, I will crawl", said the old man.				

2.10.3.11. Borrowing (BOR)

The Borrowing tier (**BOR**) contains the annotation of borrowed lexical items. Both the origin of the item in question and the type of borrowing are 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 Dolgan, there are Russian (RUS), Evenki (EV), Mongolic (MONG) and Nganasan (NGAN) borrowings. In the case of uncertainty, a per cent sign is added before the tag, as in (26). For the type of borrowing, the following tags are used (cf. also Arkhipov (2020: Ch.5)).

Table 14: 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

(24)

ref	AsKS_19XX_Amulet_nar.206 (001.203)

tx	Oru͡o	ikki	ma:mɨtɨ	halga:bɨt.
mb	Oru͡o	ikki	ma:mɨt-ɨ	halga:-bɨt
mp	Oru͡o	ikki	ma:bɨt-nI	halga:-BIT
ge	Oru͡o.[NOM]	two	noose.for.catching.reindeer-ACC	combine-PST2.[3SG]
BOR			EV:cult	
fe	Oru͡o combined two nooses for catching reindeer.			

(25)

ref	AsKS_19XX_Amulet_nar.261 (001.257)				
tx	“U͡olbar	kuppun	bɪe͡rɪe͡kpin	na:da”,	di:r.
mb	u͡ol-ba-r	kup-pu-n	bɪe͡r-ɪe͡k-pi-n	na:da	d-i:r
mp	u͡ol-BA-r	kut-BI-n	bɪe͡r-IAK-BI-n	na:da	di͡e-I:r
ge	son-1SG-DAT/LOC	amulet-1SG-ACC	give-PTCP.FUT-1SG-ACC	need.to	say-PRS.[3SG]
BOR				RUS:mod	
fe	“I have to give my amulet to my son”, he thinks.				

In the case of grammatical borrowings, the structure of the tag is generally the same, whereby only MONG (Mongolic) and EV (Evenki) are relevant here. However, the relevant morpheme’s gloss is also included in the tag, written in round brackets. This is, as can be seen in (26), to identify the morpheme to which the borrowing tag refers.

(26)

ref	SuON_KuNS_19990303_HardLife_conv.SuON.050 (001.067)		
tx	Ogonn’orum	kepse:čči	ete.
mb	ogonn’or-u-m	kepse:-čči	e-t-e
mp	ogonn’or-I-m	kepse:-A:ččI	e-TI-tA
ge	husband-EP-1SG	tell-PTCP.HAB	be-PST1-3SG
BOR	%MONG:cult	MONG:gram (PTCP.HAB)	
fe	My husband used to tell [it].		

2.10.3.12. 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 15: 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 16: 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 charts show some examples of annotating both borrowing phonology and borrowing morphology:

(27)

ref	MiXS_1967_SoldierInSecondWorldWar_nar.015 (001.015)		
tx	[…] kihi	gojobu:n	buolbutun, […]
mb	kihi	gojobu:n	buol-but-u-n
mp	kihi	gojobu:n	buol-BIT-tI-n
ge	human.being.[NOM]	wound.[NOM]	be-PTCP.PST-3SG-ACC
BOR		EV:core	
BOR-Phon		fortition	
BOR-Morph		dir:bare	
fe	[…] that people were wounded, […]		

The Evenki original is *gojowu:n* (Stachowski 1993: 86); hence, the approximant *w* is strengthened to the plosive *b*, therefore "fortition" in the BOR-Phon tier. As it is inserted without loanword morphology and there is no further inflection, there is the tag "dir:bare" in the BOR-Morph tier.

(28)

ref	MiXS_1967_SoldierInSecondWorldWar_nar.051 (001.051)			
tx	[…] n'em'ester	samal'ottara	kötön	kele-kele […]
mb	n'em'es-ter	samal'ot-tara	köt-ön	kel-e-kel-e
mp	n'em'ec-LAr	samal'ot-LArA	köt-An	kel-A-kel-A
ge	German-PL.[NOM]	airplane-3PL.[NOM]	fly-CVB.SEQ	come-CVB.SIM-come-CVB.SIM
BOR	RUS:cult	RUS:cult		
BOR-Phon	lenition			
BOR-Morph	dir:infl	dir:infl		
fe	[…] as the airplanes of the Germans were flying coming, […]			

As the Russian original is *n'em'ec*, the final affricate *ts* is weakened to the sibilant *s*, so the respective tag is "lenition" in BOR-Phon. As both borrowings are inserted without loanword morphology and are further inflected, the tags in BOR-Morph are "dir:infl".

2.10.3.13. 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 the languages of the code switch and the

type of the code switch are annotated, namely according to the scheme <LANGUAGE:type>. The language is mostly Russian (RUS), and some instances of Evenki (EV) are also found. The tag set for the type of code switch is the following:

Table 17: 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 chart shows an example of annotating code-switching:

(29)

ref	AkNN_KuNS_2002_LifeHandicraft_conv.031 (001.038)					
tx	D‿ie,	kvart'i:ru	dal'i	vs'o	tako:je,	d'e.
mb	d'‿ie					d'e
mp	d'‿ie					d'e
ge	house.[NOM]					well
CS		RUS:int.ins				
fe	A house, an apartment they gave and stuff like that.					

2.10.3.14. Existential, locative and possessive predication (ExLocPoss)

The ExLocPoss tier provides the annotation of existential, locative and possessive predication. Existential and locative predications express the temporary presence/absence of a referent X (figure) at a place Y (ground). In locative predication, the figure serves as the starting point for the perspectivization of the state of affairs, whereas it is the ground in existential predication. As a corollary, the figure is prototypically definite and topical in locative predication, whereas it is prototypically indefinite in existential predication, belonging to the focus domain. Possessive predications express that one referent Y (possessee) belongs to another referent X (possessor); prototypically, this relationship is again temporary, and the possessor has control over the possessee. In the case of inalienable possession (mostly, kinship and body terms), the latter does not hold. The core structures are thus the translational equivalents of:

"At Y, there is (no) X" (existential)

"X is (not) at Y" (locative)

"X has (no) Y" (possessive)

Koch's (2012) *generic existentials* and Haspelmath's (2022) *hyparctic clauses* (e.g. English *there are (many) unhappy people*; *there is a/no God*) are subsumed under existential clauses in the annotation. Inverted possessive clauses, i.e. *appertentive clauses* in Haspelmath's (2022) terms, (e.g. English *the book is John's*) are subsumed under possessive clauses in the annotation.

The annotation scheme includes the three functional domains *existential* (Ex), *locative* (Loc) and *possessive* (Poss), the coding strategy (see below), as well as the polarity (Aff or Neg) of the clause. The annotation tags have the format **Domain:Strategy.Polarity**. Table 18 lists the tags used for annotating existential, locative and possessive predication.

Table 18: Tags for annotating existential, locative and possessive predication

Tag	Comment
Functional domain	
Ex	existential predication
Loc	locative predication
Poss	possessive predication
Coding strategy and polarity	
:Zero.Aff	no lexical linking element, does not exclude pn-marking of figure/possessor at ground/possessee; affirmative

:Zero.Neg	no lexical linking element, does not exclude pn-marking of figure/possessor at ground/possessee; negative
:Cop.Aff	copula as linking element (*e-* and *buol-*); affirmative
:Cop.Neg	copula as linking element (*e-* and *buol-*); negative
:Ex.Aff	affirmative existential item (*ba:r*) as linking element
:Ex.Neg	negative existential item (*huok*) as linking element
:PosV.Aff	posture verb (e.g. *tur-* 'stand') as linking element; affirmative
:PosV.Neg	posture verb (e.g. *tur-* 'stand') as linking element; negative

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

(30)

ref	AnIM_AnMSp_2009_Holiday_conv.AnMSp.010 (001.016)				
tx	Onno	emie	kɨ:llar	ba:llar	[...].
mb	onno	emie	kɨ:l-lar	ba:l-lar	
mp	onno	emie	kɨ:l-Lar	ba:r-Lar	
ge	there	again	wild.reindeer-PL.[NOM]	EX-3PL	
ExLocPoss	Ex:Ex.Aff				
fe	There are again wild reindeer, [we wounded one but did not find it].				

(31)

ref	PoS_PrG_1964_Lyybyra_flk.080				
tx	[...]	Lɨ:bɨra	kanna	da	huok.
mb		Lɨ:bɨra	kanna	da	huok
mp		Lɨ:bɨra	kanna	da:	huok
ge		Lyybrya.[NOM]	where	INDEF	NEG.EX
ExLocPoss		Loc:Ex.Neg			
fe	[He looked into the chimney,] Lyybyra is nowhere.				

(32)

ref	KiES_KiLS_2009_Life_nar.KiES.040 (001.047)					
tx	[...]	onton	ol	agɨs	ogolo:kpun	bejem.
mb		onton	ol	agɨs	ogo-lo:k-pun	beje-m
mp		onton	ol	agɨs	ogo-LA:K-bIn	beje-m
ge		then	that	eight	child-PROPR-1SG	self-1SG.[NOM]
ExLocPoss		Poss:Zero.Aff				
fe	[My mother is old, my elder brother is old,] then I have eight children myself.					

2.10.3.15. Free translation (fe, fg, fr)

The free translation tiers (**fe**, **fg** and **fr**) give a 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 Arkhipov (2020: Ch. 3). The following chart shows an example:

(33)

ref	AsKS_19XX_Amulet_nar.001 (001.001)

tx	Ihilletebit	l'it'eraturnaj	p'er'edač'anɨ.
mp	ihille:-t-A-BIT	l'it'eraturnaj	p'er'edač'a-nI
ge	listen-CAUS-PRS-1PL	literary	programme-ACC
fe	We broadcast a literary programme.		
fg	Wir übertragen ein Literaurprogramm.		
fr	Мы передаем литературную передачу.		

2.10.3.16. Literal Russian translation (ltr)

The Literal Russian translation tier (**ltr**) contains the original Russian translation of the sentence in question. In the case of the texts from [FD 2000], this means the published translation. In the case of the texts made available by the TDNT and transcribed by native speakers, the transcribers were instructed to provide a literal (sometimes word-to-word) translation, reflecting the underlying Dolgan structure. The following chart shows how the literal and the free translations may differ.

(34)

ref	AsKS_19XX_Amulet_nar.139 (001.138)						
tx	Ogonn'or	hɨŋɨl	erdegine	emɪe͡	karakta:k	bagaj	ete.
mp	ogonn'or	hɨŋɨl	er-TAK-InA	emɪe͡	karak-LA:K	bagajɨ	e-TI-tA
ge	old.man.[NOM]	young	be-COND1-3SG	also	eye-PROPR.[NOM]	very	be-PST1-3SG
fe	When the old man was young, he could see very well, too.						
fr	Старик, когда был молодой, тоже был глазастый.						
ltr	Старик молодой когда был тоже глазастый был.						

2.10.3.17. Notes (nt)

The Notes tier (nt) eventually contains notes which clarify the content of the sentence or point at something peculiar in the sentence. The notes begin with the indication of who made the note (abbreviation as listed in 2.6.8, in square brackets, followed by a colon). The following chart shows an example of it.

(35)

ref	KiMN_1975_ReindeerHerding_nar.009 (001.009)		
tx	[…] ügüs	pr'ič'ine:le:k	bu͡olar.
mp	ügüs	pr'ič'ine:-LA:K	bu͡ol-Ar
ge	many	reason-PROPR.[NOM]	be-PRS.[3SG]
fe	[As I was caring about it, when a reindeer was born], there were many reasons.		
fg	[Als ich mich früher gekümmert habe, wenn ein Rentier geboren wurde], gab es viele Gründe.		
fr	[Когда раньше я ухаживал, когда олень рождался], много причины бывало.		
nt	[DCh]: Last part of the sentence not clear, even for native speakers: There were reasons for what?		

References

Arkhipov, Alexandre V. & 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. 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

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].

Chomsky, Noam. 1995. *The minimalist program.* Cambridge (Mass.): MIT Press.

Däbritz, Chris Lasse 2019. On ambiguous verb sequences in Dolgan. In: Csató, Éva Á. & Johanson, Lars & Karakoç, Birsel (eds.). *Ambiguous verb sequences in Transeurasian languages and beyond*. (Turcologica 120), 117–134. Wiesbaden: Harrassowitz.

Däbritz, Chris Lasse 2021. *Topik, Fokus und Informationsstatus: Modellierung am Material nordwestsibirischer Sprachen.* (Language, Context and Cognition 17). Berlin & Boston: De Gruyter.

Däbritz, Chris Lasse 2022. *A Grammar of Dolgan: A Northern Siberian Turkic Language of the Taimyr Peninsula*. Leiden: Brill.

FD 2000 = Efremov, Prokopij E. et al. (eds.). 2000. *Fol'klor Dolgan.* Pamyatniki fol`klora narodov Sibiri i Dal`nego Vostoka 19. Novosibirsk: Izdatel`stvo Instituta Arxeologii i Etnografii Sibirskogo Otdeleniya Rossijskoj Akademii Nauk.

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. Available online at: 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].

Haspelmath, Martin 2022. Nonverbal clause constructions. Submitted manuscript. Lingbuzz/006673. https://ling.auf.net/lingbuzz/006673 [Accessed: 17.11.2022].

Johanson, Lars. 2021. *Turkic.* Cambridge: Cambridge University Press.

Junghanns, Uwe & Zybatow, Gerhild. 2009. Grammatik und Informationsstruktur. In: Gutschmidt, Karl et al. (eds.). *Die slavischen Sprachen.* Handbücher zur Sprach- und Kommunikationswissenschaft. Bd. 32, 2. Berlin: De Gruyter, 684–707.

Koch, Peter 2012. Location, existence, and possession: A constructional-typological exploration. *Linguistics* 50(3), 533–603.

Stachowski, Marek. 1993. *Dolganischer Wortschatz.* Prace językoznawcze 114. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Ubryatova, Elizaveta I. 1985. *Yazy`k noril`skix dolgan.* Novosibirsk: Nauka.

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. Available online at: http://www.iskolakultura.hu/index.php/wpcl/article/view/10611/10503. [Accessed: 19.08.2019].

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

Label	Meaning
1	first person
2	second person
3	third person
ABL	ablative case
ACC	accusative case
ADJZ	adjectivizer
ADM	admonitive mood
ADVZ	adverbializer
AFF	affirmative
AGN	agent noun
APRX	approximative
ASSOC	associative
ASST	assistive
CAUS	causative
COLL	collective
COMP	comparative case
COND1	conditional mood 1
COND2	conditional mood 2
CVB.ANT	anterior converb
CVB.COND	conditional converb
CVB.PURP	purposive converb
CVB.SEQ	sequential converb
CVB.SIM	simultaneous converb
DAT/LOC	dative-locative case
DIM	diminutive
DISTR	distributive (numerals)
DRV	(unknown) derivational suffix
DU	dual
EMOT	emotive
EMPH	emphasis
EP	epenthetic vowel
EVID	evidential (particle)
EX	existential
EXCL	exclamative (particle)
FREQ	frequentative
FUT	future tense
GEN	genitive case
HAB	habitual
IMP	imperative mood
INCH	inchoative
INDEF	indefinite
INFER	inferential
INSTR	instrumental
INTJ	interjection
INTL	intentional
INTNS	intensive/intensifier
ITER	iterative
LIM	limitative
MID	middle
MLTP	multiplicative (numerals)
MOD	modal (particle)
MOM	momentaneous
MULT	multiplicative (actionality)

NEC	necessitative mood
NEG	negation
NEG.CVB	negative converb
NEG.CVB.SIM	negative simultaneous converb
NEG.EX	existential negation
NEG.PTCP.PRS	negative present participle
NEG.PTCP.PST	negative past participle
NMLZ	nominalizer
NOM	nominative
ONOM	onomatopoetic
ORD	ordinal numeral
PART	partitive
PASS	passive
PFV	perfective
PHIL	philative (i.e., ...phile)
PL	plural
POSS	possessive
POT	potential mood
PROB	probabilitive mood
PROPR	proprietive
PRS	present tense
PST1	past tense 1
PST2	past tense 2
PTCL	particle
PTCP.COND	conditional participle
PTCP.FUT	future participle
PTCP.HAB	habitual participle
PTCP.PRS	present participle
PTCP.PST	past participle
Q	interrogative (particle/clitic)
RECP/COLL	reciprocal-collective
REFL	reflexive
SEM	semelfactive
SG	singular
SIM	similative
TR	transitivizer
VBZ	verbalizer

Appendix 2. Dolgan bound morphemes in alphabetical order

Marker	Abbreviation	Function
-^0	NOM	nominative case
-^0	3SG	third person singular
-^0	IMP.2SG	imperative mood, second person singular
-A	EP	epenthetic vowel
-A	PRS	present tense
	CVB.SIM	simultaneous converb
-A:	VBZ	verbalizer
-A:ччI	HAB	habitual aspect
	PTCP.HAB	habitual participle
-A:jA	ADM	admonitive mood
	POT	potential mood
-A:ktA:	FREQ	frequentative
	EMOT	emotive
-A:r	FUT	future tense
-A:rAj	ADM.3SG	admonitive mood, third person singular
	POT.3SG	potential mood, third person singular
-A:rI	CVB.PURP	purposive converb
-A:T	CVB.ANT	anterior converb
-AlA:	FREQ	frequentative
-An	CVB.SEQ	sequential converb
-Ar	PRS	present tense
	PTCP.PRS	present participle
-AttA:	MULT	multiplicative
-BA	1SG	first-person singular possessive suffix (in dative-locative case of possessive declension)
-BA	NEG	negation
-BAkkA	NEG.CVB.SIM	negative simultaneous converb
-BAktA:	INCH	inchoative
-BAT	NEG.PRS	negative of present tense
	NEG.PTCP.PRS	negative present participle
-BAtAK	NEG.PST2	negation of past tense 2
	NEG.PTCP.PST	negative past participle
-BI	1SG	first-person singular possessive suffix (in possessive declension)
-BIччA	CVB.COND	conditional converb
-BIn	1SG	first person singular (predicative ending set)
-BIt	1PL	first-person plural possessive suffix (in nominative)
		first person plural (predicative ending set)
		first person plural (possessive ending set)
-BIT	PST2	past tense 2
	PTCP.PST	past participle
-BItI	1PL	first-person plural possessive suffix (in possessive declension)
-čA:n	DIM	diminutive
-čAk	ADVZ	adverbializer
	DIM	diminutive
-ччA	APRX	approximative
-ччI	ADVZ	adverbializer
-čI	INCH	inchoative
-ČIt	AGN	agent noun
-GA	DAT/LOC	dative-locative case
-GA	2SG	second-person singular possessive suffix (in dative-locative of possessive declension)
-GAr	DAT/LOC	dative-locative case (in possessive declension)
-GI	2SG	second-person singular possessive suffix (in possessive declension)

-GI	ADJZ	adjectivizer
-GIn	2SG	second person singular (predicative ending set)
-GIt	2PL	second-person plural possessive suffix (in nominative)
		second person plural (predicative ending set)
		second person plural (possessive ending set)
-GItI	2PL	second-person plural possessive suffix (in possessive declension)
h-	EMPH	emphasis
-I	EP	epenthetic vowel
-I	ADVZ	adverbializer
-I:	ADJZ	adjectivizer
-I:	NMNZ	nominalizer
-I:	CVB.SIM	simultaneous converb
	PRS	present tense
-I:hI	PROB	probabilative mood
-I:m	IMP.1SG	imperative mood, first person singular
-I:r	PRS	present tense
	PTCP.PRS	present participle
-IAgIŋ	IMP.1PL.IN	imperative mood, first person plural
-IAjAk	LIM	limitative numeral
-IAk	IMP.1DU	imperative mood, first person dual
-IAK	FUT	future tense
	PTCP.FUT	future participle
-IAktI	INTL	intentional
-IAktIn	INTL.3SG	intentional, third person singular
-IAktInnAr	INTL.3PL	intentional, third person plural
-IAlA:	FREQ	frequentative
-IAn	COLL	collective numeral
-IAr	CAUS	causative
=Ij	Q	interrogative clitic
-ijek	DRV	(unknown) derivational suffix
-In	ADVZ	adverbializer
-In	3SG	third person singular (at anterior converb)
-InA	3SG	third person singular (conditional ending set)
-InnAr	CAUS	causative
-Is	ORD	ordinal numeral
-ItAlA:	FREQ	frequentative
-j	VBZ	verbalizer
-k	NMLZ	nominalizer
	ADJZ	adjectivizer
-k	DRV	(unknown) derivational suffix
-kA	DIM	diminutive
	INTNS	intensive/intensifier
-kA:N	DIM	diminutive
	INTNS	intensive/intensifier
	LIM	limitative
-kAj	ADJZ	adjectivizer
-ke:če:n	INTNS	intensive/intensifier
-kɨn	INTNS	intensive/intensifier
-kInA	2SG	second person singular (conditional ending set)
-kItInA	2PL	second person plural (conditional ending set)
-ku:	EMPH	emphasis
-LA:	VBZ	verbalizer[17]
-LA:gI	ADJZ	adjectivizer

[17] -LA: is also used to integrate borrowed verbs from Russian into Dolgan.

-LA:K	PROPR	proprietive
	NEC	necessitative mood
-LAN	REFL/MID	reflexive/middle
	VBZ	verbalizer
-LAr	PL	plural
-LAr	3PL	third person plural (predicative ending set)
-LArA	3PL	third-person plural possessive suffix (nominative)
		third person plural (possessive ending set)
-LArI	3PL	third-person plural possessive suffix (in possessive declension)
-LI:	SIM	similative
	DISTR	distributive numeral
-LI:N	ASSOC	associative
-LIk	NMLZ	nominalizer
	ADVZ	adverbializer
-lIN	PASS/REFL	passive/reflexive
-m	1SG	first-person singular possessive suffix (nominative)
		first person singular (possessive ending set)
-m	NEG	negation
-mInA	NEG.CVB	negative converb
-msAk	ADJZ	adjectivizer
	PHIL	philative (i.e., …phile)
-mslj	MOM	momentaneous
	DRV	(unknown) derivational suffix
-n	ACC	accusative case (in possessive declension)
-n	GEN	genitive case (in possessive declension)
-n	REFL	reflexive
	MID	middle
-n	VBZ	verbalizer
-nA	PART	partitive case (in possessive declension)
-nAn	INSTR	instrumental case
-nI	ACC	accusative case
-ŋ	2SG	second-person singular possessive suffix (nominative)
		second person singular (possessive row)
-ŋ	IMP.2PL	imperative mood, second person plural
-ŋnA:	ITER	iterative
	SEM	semelfactive
	VBZ	verbalizer
-plnA	1SG	first person singular (conditional ending set)
-pltlnA	1PL	first person plural (conditional ending set)
-r	DAT/LOC	dative-locative case (in possessive declension)
-r	CAUS	causative
	INTNS	intensive/intensifier
-s	RECP/COLL	reciprocal-collective
	MID	middle
	ASST	assistive
-s	NMLZ	nominalizer
-skA	NMLZ	nominalizer
-t	CAUS	causative
	INTNS	intensive/intensifier
	MID	middle
	PASS	passive
	TR	transitivizer
-tA	3SG	third-person singular possessive suffix (nominative)
		third person singular (possessive ending set)
	POSS	possessive

-TA	PART	partitive case
-TA	MLTP	multiplicative numeral
-TA:	ITER	iterative
-TA:gAr	COMP	comparative case
-tAj	INTNS	intensive/intensifier
-TAK	PTCP.COND	conditional participle
	COND1	conditional mood 1
	INFER	inferential
-TAr	CAUS	causative
	PASS	passive
-TAR	COND2	conditional mood 2
-tArInA	3PL	third person plural (conditional ending set)
-tI	3SG	third-person singular possessive suffix (possessive declension)
-TI	PST1	past tense 1
-TIj	INCH	inchoative
-TIn	IMP.3SG	imperative mood, third person singular
-TInnAr	IMP.3PL	imperative mood, third person plural
-ttAn	ABL	ablative case