



INEL Kamas corpus 2.0

User documentation

Alexandre Arkhipov, Chris Lasse Däbritz, Valentin Gusev

1. Introduction

1.1. Objective of the corpus

The present corpus of the Kamas language has been created as part of the long-term research project INEL (“Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages”), whose primary goal is to create digital annotated corpora of several languages of Northern Eurasia, making possible typologically aware corpus-based grammatical research.

The INEL Kamas corpus aims to bring together the whole body of available recorded spoken Kamas data. This is the first publicly available annotated and searchable digital resource for Kamas texts ever.

1.2. Kamas language

1.2.1. Description

Kamas belongs to the Samoyedic branch of the Uralic language family. All the main sources document Forest Kamas varieties spoken in the settlement of Abalakovo, in the present Krasnoyarsk Krai in Southern Siberia, on the northern slopes of the East Sayan Mountains. The language became extinct by the late XXth century with the death of its last known speaker, Klavdiya Plotnikova, in 1989. It was documented by several researchers in XVIII–XX centuries, however, substantial Kamas texts were only recorded in the XXth century.

Kamas is an agglutinating language with some flective elements, predominantly suffixing. Nouns have two declension types, absolute and possessive declension. In verbal morphology subjective and objective conjugation are distinguished, the latter as an incomplete paradigm. Kamas is rich in derivational suffixes, especially in verbal aspect markers. Its phonology and lexicon, as well as such grammatical features as converb constructions and aspectual auxiliaries, show impact of contact with Turkic languages. Part of the Kamas speakers eventually shifted to a local Turkic variety (Kacha dialect of Khakas), while the others finally shifted to Russian.

“Pre-shift” vs. “post-shift” Kamas

The present corpus consists of two parts: folklore texts collected by Kai Donner in 1912–1914, before the language shift, and transcribed audio recordings from two last (semi-)speakers made

between 1964 and 1970. The latter represent what was left of the Kamas language after the community shifted to Russian completely. These two collections will be referred to as Donner's ("pre-shift") collection and the "post-shift" collection. They differ both in the form of the source material and in the nature of the language data.

Kai Donner worked with Kamas speakers in 1912 and 1914. His posthumously edited Kamas fieldwork materials ("*Kai Donners Kamassisches Wörterbuch...*") [Joki 1944] include a dictionary, a grammar sketch and several exemplary texts, mostly folklore. The text collection is written down in detailed phonetic transcription (FUT). Donner also had collected sound recordings, using phonographic wax cylinders. The original cylinders are now lost, and only a small part (ca. 8 minutes) has been recorded on tape and later digitized (see Klumpp 2013). It contains versions of some texts from the written material but not the exact same texts; these recordings are not yet fully transcribed.

According to Donner, the language proficiency was already declining and the language was not as rich and elaborate as it used to be; already at that time only some people over 45 or 50 spoke the language well. Yet Donner's collection is the oldest collection of coherent Kamas texts.

The last Kamas speaker, Klavdiya Zakharovna Plotnikova (Andzhigatova), was discovered during a toponymic fieldtrip by A. Matveyev and his students in 1963. By that time she had not spoken her language for some 20 years. However, she still remembered the language to some extent, and in the subsequent years reactivated its use while working with linguists. She was recorded on tape by several researchers in 1964–1970. There exist two tapes with recordings of the other speaker, Aleksandra Eliseevna Semyonova (Dzhibyeva), taken around 1964 in Krasnoyarsk; she died soon afterwards. Both Plotnikova and Semyonova had most probably not fully acquired Kamas in their childhood, and were recorded after many years of not using the language, and are better regarded as language rememberers. Semyonova (and probably Plotnikova too) also spoke Khakas (Kacha dialect) to some extent. The language in the post-shift corpus is a product of intensive erosion, with very restricted grammar and lexicon, inconsistent use of grammatical markers and speech disfluencies, and effects of Russian influence on all language levels.

1.2.2. Language Codes

ISO-639-3 code: **xas**

Glottolog code: **kama1378**

1.2.3. Dialectal subdivisions

Kamas probably had three dialects, the Koybal (documented by S. P. Pallas in XVIII and by G. Spasski in early XIX, assimilated by Turkic populations and extinct in the late XIXth century), the Forest Kamas, which includes the most substantially documented varieties, and the least known Steppe Kamas (documented in XVIII).

The Forest Kamas spoken in Abalakovo, from where all the available recorded text originate, was not internally homogeneous. M. A. Castrén documented two different varieties, Forest I and Forest II, in 1847. Kai Donner documented two main varieties belonging to two Kamas tribes or clans, the Fat clan (Andzhigatov family) and the Eagle clan (Ashpurov family), the latter variety most reliably identifiable with Castrén's Forest I. Donner's materials also contain some variants which are not consistent with either of the Fat and Eagle varieties but still presumably closer to Forest I than to Forest II.

1.3. Archiving

The corpus comprises source media files (whenever available) along with the annotated transcripts in *EXMARaLDA*¹ transcript formats and metadata descriptions in *EXMARaLDA Coma*² format (see 2.7 and 2.9 for details).

The corpus is archived and published by the Research Data Repository of the Universität Hamburg³ under open-access conditions with Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0).⁴

1.4. Citation

The corpus is to be cited as follows:

Gusev, Valentin; Klooster, Tiina; Wagner-Nagy, Beáta. 2023. "INEL Kamas Corpus." Version 2.0. Publication date 2023-12-31. <http://hdl.handle.net/11022/0000-0007-FC25-4>. 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 Kamas corpus has been created as part of the long-term INEL project ("Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages"), 2016–2033. For an overview of the project, see [Arkhipov, Däbritz 2018]. The Kamas subproject spanned three years from January 2016 to December 2018. Version 1.0 was published in 2019. This maintenance release was prepared in 2023.

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://inel.corpora.uni-hamburg.de/>.

Project leader

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

Researchers

Dr. Alexandre Arkhipov, Research coordinator

Tiina Klooster, M.A. (February 2016 – September 2017)

Dr. Valentin Gusev (November 2017 – December 2018)

Josefina Budzisch, M.A.

Chris Lasse Däbritz, M.A.

Hannah Wegener, M.A.

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

¹ <http://exmaralda.org/en/>, last access: 20.12.2023

² <http://exmaralda.org/en/corpus-manager-en/>, last access: 20.12.2023

³ <https://www.fdr.uni-hamburg.de/communities/inel/>, last access: 20.12.2023.

⁴ <https://creativecommons.org/licenses/by-nc-sa/4.0/>, last access: 20.12.2023.

Developers

Timm Lehmborg, M.A., Technical coordinator (January 2016 – January 2023)

Daniel Jettka, M.A. (January 2016 – February 2021)

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

Anne Ferger, M.A. (April 2017 – March 2021)

Elena Lazarenko (since May 2021)

Aleksandr Riaposov (since April 2021)

Student assistants

Hannes Klitzing (September – December 2016)

Olesya Degtyareva (October 2016 – December 2017)

Felix Templin (April 2016 – June 2018)

Ozan Özdemir (August 2018 – December 2019)

Felicitas Otte (April 2019 – December 2019)

1.6. Acknowledgements

Funding

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

Data sources

Texts collected by Kai Donner come from an unpublished edition of Kai Donner's manuscripts by Prof. Gerson Klumpp (University of Tartu). The manuscripts were first philologically edited by Hartmut Katz and Gerson Klumpp. The resulting version was enhanced, rendered into a phonological transcription and morphologically glossed by Gerson Klumpp. It is this edition of the texts which was adapted for INEL by Tiina Klooster and further edited by Valentin Gusev. The German translation by A. Joki are taken from the published version of the texts [Joki 1944]. The phonetic transcription based on Klumpp's edition of Donner's manuscripts is now provided in the stl tier.

Prof. Klumpp has also provided extensive consultations on Kamas during the project. The phonological transcription used in the INEL Kamas corpus follows in most respects the transcription system developed by him, although some changes were made to bring it closer to transcriptions in other INEL corpora. Parts of transcription of the sound files were done by Tiina Klooster in close cooperation with Prof. Klumpp. Partial transcriptions from their earlier project (parts of AEDKL tapes SU0211, SU0222, SU0223, SU0225-SU0229) served as basis for transcriptions of these texts done in INEL.

Archive of Estonian Dialects and Kindred Languages of the University of Tartu, Estonia (AEDKL, or TÜEMSA; <https://murdearhiiv.ut.ee/>) provided the recordings of Kamas speech made by Ago Künnap in Abalakovo and by Tiit-Rein Viitso in Tartu, as well as the digitized fragment of the surviving copy of Kai Donner's phonograph recording.

The Institute for the Languages of Finland archive, Helsinki, Finland (KOTUS; <https://www.kotus.fi/>) provided the recordings of Klavdiya Plotnikova made by Jaakko Yli-Paavola in Tallinn in 1970.

Scanned pages from the *Kai Donners Kamassisches Wörterbuch* [Joki 1944] containing texts collected by Kai Donner published online courtesy of the Finno-Ugrian Society (<https://www.sgr.fi/>).

Part of translations into German were done by Ralph Reindler.

1.7. New in release 2.0

- In texts from Donner's collection, phonetic transcription according to Klumpp's edition of Donner's manuscripts has been added (as stl tier)
- Five texts which were originally split between different tapes have been merged, as well as respective parts of recordings. Sentences in each resulting text are numbered throughout
 - PKZ_196X_Alenushka_flk + PKZ_196X_Alenushka_continuation_flk > PKZ_196X_Alenushka_flk
 - End of PKZ_196X_SU0226 starting from PKZ_196X_SU0226.203 (210) + PKZ_196X_Alenushka2_continuation_flk > PKZ_196X_Alenushka2_flk
 - PKZ_196X_BlacksmithAndMerchant_flk + PKZ_196X_BlacksmithAndMerchant_cont_flk > PKZ_196X_BlacksmithAndMerchant_flk
 - PKZ_196X_Finist_flk + PKZ_196X_Finist_continuation_flk > PKZ_196X_Finist_flk
 - PKZ_196X_StupidWolf_flk + PKZ_196X_StupidWolf_continuation_flk > PKZ_196X_StupidWolf_flk
- Part of the texts are now annotated for existential, locative and possessive predication (ExLocPoss tier, by C.L. Däbritz)
- Numerous corrections in glosses, other annotations and transcriptions, including:
 - Fuller and more consistent transcription, glossing and annotations of borrowings
 - Vowel length is marked in mp tier in *ba:zo?* 'again', *bü:z'e* 'man' and *sa:gər* 'black'
 - Corrections in disambiguation of polysemous or homonymous morphemes:
-zi? "INS"/"COM", -də "LAT"/"3SG", mo- "can/become/want | мочь/стать/хотеть"
 - Possessive suffix unmarked for case: "NOM/GEN/ACC" > "POSS"
 - Glosses for personal pronouns were changed to uniform labels: "I | я" > "PRO1SG", "we | мы" > "PRO1PL", "you | ты" > "PRO2SG", "you.PL | вы" > "PRO2PL"
 - Fuller annotations of code-switching and calques (CS tier)
- Added ELAN *.eaf as a supplementary end-user file format for all transcripts

2. The corpus

2.1. The language(s) of the corpus

2.1.1. Content

The principal language of content in the corpus is Kamas. In the post-shift audio collection, there are also prompts from the researchers in Russian and a considerable amount of Russian code-switching in the speech of Klavdiya Plotnikova. Occasionally there are some remarks in Estonian or Finnish by the researchers. Finally, there are a dozen utterances in Khakas in the recordings of Aleksandra Semyonova.

There is always only one main transcription tier (per speaker), using the common INEL transcription style (see 2.10.2).

2.1.2. Annotations

The main annotation language in the corpus is English.

The main content transcript is translated into English, Russian and German (see tiers **fe**, **fr**, **fg**). For texts from the Donner's collection, the original translation into German is given in tier **ltg** as provided in [Joki 1944]; it is a somewhat archaic and non-standard form of German supplied by A. Joki, the editor of Donner's materials, which served as basis for the other translations provided.

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

2.1.3. Metadata

The language of metadata is English. Russian spellings of the personal names and place names are also provided in communications and speaker metadata. English spellings of Russian names in the metadata follow the GOST 7.79 System B transliteration standard.

2.2. Sources

The corpus consists of two parts: an unpublished edition of texts from Kai Donner's collection (see 1.6), and transcribed audio recordings from two last (semi-)speakers, Klavdiya Plotnikova (most of data) and Aleksandra Semyonova, made between 1964 and 1970.

Audio recordings by Aleksandr Matveyev and Ago Künnap were made during the 1960s. Deposited in the Archives of Estonian Dialects and Kindred Languages of the University of Tartu. Recorded on tapes, digitized in 2010.

Audio recordings by various Estonian and Finnish researchers were made in 1970. Deposited in the KOTUS archive in Helsinki. Recorded on tapes, digitized in 2006.

2.3. Content

The Donner's collection contains 16 texts, mostly folklore, with the addition of two sets of riddles, two short prayers (given here as one text) and a lamentation song. The latter is classified here as "song", all the other texts as "folklore".

Recordings of Klavdiya Plotnikova present a mixture of genres. These include folklore texts, including retelling the texts from Donner's collection and retelling Russian folk tales; interviews/conversations on the life of Plotnikova and of the Kamas people in general; daily life stories; translations of individual sentences from Russian prompts. Except for the longer folklore texts, the tape is often restarted, usually omitting the researcher's prompts and pauses, and many recordings are thus highly fragmented (often with some 40 to 80 fragments per tape). This is why many Plotnikova recordings are classified here as "miscellaneous" and treated a single communication (text), while longer consistent fragments are usually extracted and attributed to the genres of "folklore" or "narrative".

2.4. Selection

The present corpus covers all Kamas texts represented in [Joki 1944] and all the Kamas spoken data preserved at AEDKL (TÜEMSA) and at the KOTUS archive, with the exception of Kai Donner's phonographic recording (AEDKL item SU0233) which is included without a transcript.

2.5. Corpus size

This version of the corpus contains 154 transcripts (98 folklore, 1 song, 11 narratives, 44 miscellaneous) of 4 known speakers (for 5 texts the speaker is unknown) with 13 786 sentences and 63 810 words including Russian speech of the interviewers, Russian code-switching and speech disfluencies of Klavdiya Plotnikova. With Russian fragments and false starts excluded, the word count is approximately 46,500 in the post-shift part and 2,500 in the pre-shift part, i.e. 49,000 in total. The overall duration of the transcribed audio data is ca. 14 h.

2.6. Naming Conventions

2.6.1. Name of the corpus

The name of the corpus is "INEL Kamas Corpus".

2.6.2. Orthography conventions in the corpus

The Kamas transcription used in the present corpus is an adaptation of the phonemic transcription developed by Prof. Gerson Klumpp. It is based on a simplified and standardized Finno-Ugric Transcription (FUT). The list of occurring phonemes and their representing characters is enclosed as [Appendix 1](#). For the phonetic transcription of Donner's texts, see 2.10.2.2.

Long vowels are marked by the character : (triangular colon) following the vowel sound (e.g. *ša:škən* 'magpie'). The existence of inherent vowel length distinctions in Kamas has been a matter of uncertainty. In previous versions of the corpus, only long vowels which have emerged as a result of contraction inside the stem were marked as such. However, in the current version long vowels are also marked (in mp tier) in three roots where the length appears to be lexically specified: *ba:zo?* 'again', *bü:z'e* 'man' and *sa:gər* 'black'. In Donner's texts, long vowels are marked with double characters in a few cases where they are a result of compensational lengthening on the morpheme boundary.

Palatalization is marked by the symbol ' (apostrophe) following the consonant (e.g. *t'äga* 'river').

Since there is a lot of individual variation in pronunciation, the phonemic transcription brings the word forms back to their underlying form. There can be a noticeable difference between the phonetic realization of a word and the phonemic form.

2.6.3. Folder structure

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

Folders with text transcripts, organized by genre:

- "flk" (folklore texts)
- "nar" (narrative texts)
- "misc" ("miscellaneous")
- "song" (song)

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

- annotated transcript in EXMARaLDA EXB and EXS formats (*.exb, *.exs)

- scanned pages from the [Joki 1944] publication, in PDF (*.pdf) (for texts from Donner's collection)
- sound file in WAV (*.wav) (for texts with audio source)
- annotated transcript converted to ELAN EAF format (*.eaf)
- annotated transcript converted to ISO/TEI format (*.xml)

Supplementary folders:

- "documentation" (user documentation)

Individual files:

- "kamas.coma" (main metadata file)

2.6.4. Transcripts

The names of the transcript files have the structure `Speaker_DateOfRecording_Title_Genre`, the same as the respective communication code in the metadata (see 2.6.6.1 for details). The segmented transcript files additionally have a "_s" suffix in the end of their name. The file extensions are .exb and .exs for the basic and segmented transcript files respectively (see 2.7.1).

2.6.5. Media

The names of the audio files (for the post-shift collection) have the structure `Speaker_DateOfRecording_Title_Genre`, the same as the respective communication code in the metadata (see 2.6.6.1 for details). The same goes for the scanned pages of the *Kamassisches Wörterbuch* [Joki 1944] (for Donner's collection) in PDF format.

2.6.6. Metadata

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

2.6.6.1. Names of communications

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

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

The communication short title is a (possibly shortened) version of the English title, spelled without spaces, dashes or other non-letter characters, with all initial capitals. This English title is usually a translation of the Russian title, which is generally given by the corpus creators. For the tapes ascribed to the "miscellaneous" genre, the title is based on the archive code of the tape (e.g. "SU0210" for AEDKL and "09340-1bz" for KOTUS). For texts parts of which were found on different archive tapes, the second part is presented separately and has a suffix "_cont" or "_continuation" appended to the title.

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

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

Name: AA_1914_Brothers_flk

Speaker: AA (see 2.6.6.2)

Date of recording: 1914

Short title: Brothers

Genre: flk (i.e. a folklore text)

2.6.6.2. Speaker codes

The codes for the speakers are made up of one letter pointing at the last name, one letter pointing at the surname and one letter pointing at the patronymic – if one of those is not known, then it is left out. E.g. AA stands for Avdakeya Andzhigatova.

AA: Andzhigatova, Avdakeya

AIN: Ashpurov, Innokentiy Nikolaevich

PKZ: Plotnikova (Andzhigatova), Klavdiya Zakharovna

SAE: Semyonova (Dzhibyeva), Aleksandra Eliseevna

NN: Unknown speaker

2.6.7. Abbreviations

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

Data collectors and editors

DK: Donner, Kai

JA: Joki, Aulis

KA: Künnap, Ago

KIG: Klumpp, Gerson

MAK: Matveev, Aleksandr

VTR: Viitso, Tiit-Rein

YPJ: Yli-Paavola, Jaakko

Project members

AAV: Arkhipov, Alexandre

BJ: Budzisch, Josefina

DCh: Däbritz, Chris Lasse

GVY: Gusev, Valentin

KIT: Klooster, Tiina

WH: Wegener, Hannah

WNB: Wagner-Nagy, Beáta

Student assistants

DO: Degtyareva, Olesya

JK: Jessen, Kis

KH: Klitzing, Hannes

OF: Otte, Felicitas

Third-party translation

ReR: Reindler, Ralph

2.7. Technical formats

2.7.1. Transcripts

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

The same content is also provided in two other formats created automatically from the EXMARaLDA transcripts.

The format of the ELAN multimedia annotator (*.eaf) is common in language documentation studies. Users of ELAN can thus directly browse the converted EAF transcripts and perform searches across the entire corpus using ELAN capabilities.

ISO/TEI (*.xml) is a TEI format recognized as an ISO standard for representation of spoken data. It is used, among other, as a source format for the Tsakorpus platform which provides online search over INEL corpora.

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”). This single file holds the metadata for the entire corpus.

2.7.3. Media

Audio files of the digitized analog tapes are provided as stored in the archives in Linear PCM WAVE format (file extension “.wav”), mono 16 bit, with sampling frequency of 44 100 Hz for files from the KOTUS archive and 48 000 Hz for files from the AEDKL (TÜEMSA) archive.

For texts from Donner’s collection, corresponding pages scanned from [Joki 1944] are provided in PDF format.

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

For Donner’s collection, the texts were converted to a simplified standard phonemic transcription and then imported into *SIL Fieldworks Language Explorer* (FLEX)⁵ for glossing.

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

For the post-shift collection, the sound files in wav format were segmented into utterances and transcribed in ELAN multimedia annotator⁶, translated into English and then imported into FLEEx for glossing.

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

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

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

When the corpus is finalized, EXMARaLDA transcripts are automatically converted into EAF and ISO/TEI formats.

2.8.2. Media

The media files provided by the archives generally did not require processing, with two exceptions. The AEDKL file SU0211 had a few seconds of reversed recording in the end, which turned out to be a fragment of an Estonian folk tale. The KOTUS file 09343:1a had a fragment recorded at double speed in the middle. Both issues are corrected in the published version.

The texts from Donner's collection were scanned from [Joki 1944] and saved in PDF format. They were also subjected to OCR to extract the German translations by A. Joki.

2.8.3. Metadata

The corpus metadata are managed by *EXMARaLDA Corpus Manager (Coma)*.

2.9. Metadata for the corpus

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

2.9.1. Naming conventions and content of the metadata

The general metadata about the whole corpus include the corpus name ("INEL Kamas 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 name which is given to the communication (see 2.6.6.1)

Description:

- **0a. Title:** Short title (in English)
- **0b. Title (RU):** Short title (in Russian)

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

- **1a. Genre:** Abbreviation of the genre of the communication: conv = conversation (also interviews), flk = folklore, nar = narrative, song = song, misc = miscellaneous
- **1b. Genres of fragments:** Communications ascribed to the “misc” genre can be further specified as containing parts classified as flk = folklore, nar = narrative, conv = conversation, sent = (separate) sentences
- **2a. Recorded by:** Abbreviation of the person by whom the communication was recorded (see 2.6.7)
- **2b. Date of recording:** Here the date of recording is given (year only)
- **3a-b. Dialect / Subdialect:** Information on the dialect used by the speaker is given here; in the case of Kamas, all the data in the corpus represent the same dialect (Forest Kamas, Forest I subdialect)
- **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 known) of transcribing
- **7a-d. Translation(s):** Code of the person who did the translation in question (Russian, English, German, source German in published edition)
- **8a. Glossed by:** Code of the person who did the morphological glossing
- **8b. Glosses checked:** Whether the glosses have been double-checked
- **9a-d. Annotation(s):** Code of the person who did the annotation in question (SeR, SyF, IST, BOR/CS; see 2.10)

Generally in the Description fields, if multiple person codes are given, esp. for translation and annotation activities, the last referred person is usually the editor of previous translations/annotations.

Location:

- **Country:** Most the communications originate from Russia, some were recorded in Estonia
- **Region:** The current administrative region is indicated
- **Settlement (LngLat):** Longitude and latitude of the place of recording
- **Settlement:** The place of recording

Languages:

- **Language code:** The ISO code of the language of communication (*xas* – Kamas; other languages are not mentioned)

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

- **1. Archive (written):** Reference to identified transcripts in the written materials of the archive of the Philological faculty of Uralic Federal University, Ekaterinburg.
- **2. Corresp. sound/written:** not used.
- **3a. Published in:** If the text has been published, entirely or in part, the publication reference is provided.
- **3b. Published in (bibtex):** BiBTeX key of the publication in the INEL bibliography.

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 taking part in a communication generally include, on the one hand, biographical information of the speaker, and on the other hand, information on his sociolinguistic background. The following information is given as exactly as possible:

Description of speaker:

- **1a-b. Family name (EN, RU)**
- **2a-b. Given name (EN, RU)**
- **3a-b. Patronymic (EN, RU)**
- **4. Tribe:** one of the two Kamas tribes in Abalakovo noted by Donner: *śil* (Fat) and *ńiga* (Eagle)
 - **5a-b. Alternate names (EN, RU):** If alternate names (e.g. maiden name, short name/diminutive) or name spellings are found, they are given here

Education: Education information is irrelevant for the Kamas corpus.

Informant of: Here the researcher is mentioned with whom the speaker worked.

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

- **1a. Ethnicity**
 - **2a-b. Ethnicity of mother / Name of mother**
 - **3a-b. Ethnicity of father / Name of father**
 - **4a-b. Ethnicity of husband/wife / Name of husband/wife**
 - **5a-b. Ethnicity of grandparents / Names of grandparents**
 - **6a-b. Family (EN, RU):** other family information

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

- **1a-b. Place of birth (EN, RU)**
- **2. Region**
- **3. Country:** Russia
- **4. Date of birth**
- **5. Date of death**
- **6a-b. Former residences (EN, RU):** If former residences prior to the work with the linguist are known, they are mentioned here
- **7a-b. Domicile:** Here the current (i.e. at the time of the recording) place of residence of the speaker is mentioned

Languages: Russian is listed as L2 for Donner's informants but as L1 for Plotnikova and Semyonova. Khakas is also listed as L2 for Semyonova. Language codes are given: *xas* — Kamas, *rus* — Russian, *kjh* — Khakas. For all Kamas speakers in the corpus, the (sub-)dialect is Forest I.

2.10. Transcription and annotation

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

2.10.1. Tier layout

Tab. 1: Overview of annotation tiers

Tier label	Tier full name	Description	Unit	Optionality
ref	Reference	Text ID + sentence number	sentence	obligatory
stl	Source text (Latin)	Donner's phonetic transcription (for texts from Donner's collection)	sentence	optional
ts	Text (sentence)	Main transcription	sentence	obligatory
tx	Text	Main transcription segmented by word	word	obligatory
mb	Morpheme breaks	Morpheme breakdown of words (morphemes dash-separated for each word)	morph	obligatory
mp	Morphophonemes (underlying)	Underlying (lexical) representation of morphemes	morph	obligatory
ge	Gloss (English)	Morpheme glosses (with lexical glosses in English)	morph	obligatory
gr	Gloss (Russian)	Morpheme glosses (with lexical glosses in Russian)	morph	obligatory
mc	Morphological category	Morphological category/part of speech for each morpheme	morph	obligatory
ps	Part of speech	Part of speech for each word	word	obligatory
SeR	Semantic role	Semantic (thematic) roles for major NPs	word / group of words	optional
SyF	Syntactic function	Syntactic functions for predicates and arguments	word / group of words	optional
IST	Information status	Information status for major NPs (given/new/accessible)	word	optional
ExLocPoss	ExLocPoss	Existential, locative and possessive predication	sentence	optional
BOR	Borrowing	Borrowings (source language and type)	word	optional
BOR-phon	Borrowing phonology	Phonological adaptations in borrowings	word	optional
BOR-morph	Borrowing morphology	Morphological adaptations in borrowings	word	optional
CS	Code switching	Code switching and calques (source language and type)	group of words	optional
fe	Free translation (English)	Free translation (English)	sentence	obligatory
fg	Free translation (German)	Free translation (German)	sentence	obligatory
fr	Free translation (Russian)	Free translation (Russian)	sentence	obligatory
ltg	Literal translation (German)	Original German translation as provided in [Joki 1944] (for texts from Donner's collection)	sentence	optional

Tier label	Tier full name	Description	Unit	Optionality
nt	Notes	Notes from corpus developers	sentence	optional

2.10.2. Transcription tiers

2.10.2.1. Main transcription tiers (tx and ts)

The main transcription tiers use the INEL Kamas transcription (see 2.6.2). The major difference between them is that **ts** presents transcriptions of entire sentences, while **tx** has the same content divided into words. The latter is the basis for the morpheme breakdown in the tier **mb** (see below) and further word-level annotations. Technically speaking, in EXMARaLDA format it is only the **tx** tier which has the type “transcription”, all other tiers being of the type “annotation”. It is thus the **tx** tier which serves as the basis for segmentation (in “segmented transcription” format, EXS), which is relevant for search using the EXAKT tool and for all sentence and word counts.

(1)

ref	AA_1914_Hare_flk.001 (001.001)	AA_1914_Hare_flk.002 (001.002)
ts	Kozan kandəbi, kandəbi.	Nugurbi toʔbdobi, nugurbinə püjebə bəppi.
fe	A hare walked and walked.	He met steppe grass, on the steppe grass he cut his nose.

(2)

ref	AA_1914_Hare_flk.001 (001.001)		
ts	Kozan kandəbi, kandəbi.		
tx	Kozan	kandəbi,	kandəbi.
fe⁷	A hare walked and walked.		

The treatment of some special cases and phenomena such as uncertainties and alternatives in transcription, unintelligible fragments, false starts and non-speech sounds is described in (Arkhipov 2020).

2.10.2.2. Donner’s transcription (stl)

For texts from Donner’s collection, his phonetic transcription is provided in the **stl** tier, following as close as technically possible the unpublished edition of Donner’s manuscripts by Gerson Klumpp (see 1.6). Note that the distinction between regular and superscript characters is not preserved.

(3)

ref	AA_1914_Hare_flk.001 (001.001)	AA_1914_Hare_flk.002 (001.002)
stl	k'õzɔn k'andəbi, k'andəbi [.]	nuyurbi toʔbtõbi, nuyurbinə p'ujëbə bə'pi.
ts	Kozan kandəbi, kandəbi.	Nugurbi toʔbdobi, nugurbinə püjebə bəppi.
fe	A hare walked and walked.	He met steppe grass, on the steppe grass he cut his nose.

⁷ “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 dot. The sentences are numbered throughout the entire text, separately for each speaker. 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*, consecutive irrespective of the speaker).

(4)

ref	AA_1914_Hare_flk.001 (001.001)
------------	--------------------------------

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. Productive derivational suffixes are segmented, while non-productive derivational suffixes are mostly not segmented and the derived stem is then glossed as a separate lexical item.

(5)

ref	AA_1914_Hare_flk.001 (001.001)		
tx	Kozan	kandəbi,	kandəbi.
mb	kozan	kandə-bi	kandə-bi
fe	A hare walked and walked.		

2.10.3.3. Morphophonemes (underlying) (mp)

The underlying morphemes tier (**mp**) shows the underlying representation of the morphs which appear separated in the **mb** tier. Stems are, hence, represented here by their lexical entry in the FLEx lexicon. Affixes are represented by their main allomorphs. All morphemes within a word are separated by hyphens. Zero morphs are not represented in this tier. For an example see Error: Reference source not found.

(6)

ref	AA_1914_Hare_flk.012 (001.012)				
tx	"Män	üjüne	jilgəndə	büjə?	bitle?bəliem."
mb	män	üjü-nə	jil-gəndə	bü-jə?	bīt-le?bə-lie-m
mp	män	üjü-nə	il-gəndə	bü-jə?	bīs-la?bə-liA-m
fe	"I'm drinking waters down at my feet."				

2.10.3.4. Gloss (ge and gr)

The gloss tiers (**ge** and **gr**) contain the English and Russian glossing of the morphemes in **mb** and **mp**. Stems receive their respective lexical glosses in the two languages, while affixes are glossed identically in latin script and mostly according to the Leipzig Glossing Rules.⁸ For the full list of glossing abbreviations, see [Appendix 2](#).

⁸ <https://www.eva.mpg.de/lingua/resources/glossing-rules.php>, last access: 22.12.2023.

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

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

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

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

(7)

ref	AA_1914_Hare_flk.012 (001.012)				
tx	"Mǎn	üjünə	jilgəndə	büjə?	bitle?bəliem."
mb	mǎn	üjü-nə	jil-gəndə	bü-jə?	bīt-le?bə-lie-m
mp	mǎn	üjü-nə	il-gəndə	bü-jə?	bīs-la?bə-liA-m
ge	I.NOM	foot-GEN.1SG	underpart-LAT/LOC.3SG	water-PL	drink-DUR-PRS-1SG
gr	я.NOM	нога-GEN.1SG	низ-LAT/LOC.3SG	вода-PL	пить-DUR-PRS-1SG
fe	"I'm drinking waters down at my feet."				

(8)

ref	AA_1914_Hare_flk.013 (001.013)			
tx	Dīgəttə	kambi	t'aktə	nükeinə.
mb	dīgəttə	ka-m-bi	t'aktə	nüke-i-nə
mp	dīgəttə	ka-n-bi	t'aktə	nüke-jə?-nə
ge	then	go-PST.[3SG]	old	woman-PL-LAT
gr	тогда	идти-PST.[3SG]	старый	женщина-PL-LAT
fe	Then he went to the old women.			

2.10.3.5. Morphological category (mc)

The **mc** tier indicates the morphological category of both lexical stems (i.e. the part of speech) and affixes (i.e. the inflectional category or the derivational process). For the list of abbreviations, see Appendix 3.

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

(9)

ref	AA_1914_Hare_flk.012 (001.012)				
tx	"Mǎn	üjünə	jilgəndə	büjə?	bitle?bəliem."
mb	mǎn	üjü-nə	jil-gəndə	bü-jə?	bīt-le?bə-lie-m
mp	mǎn	üjü-nə	il-gəndə	bü-jə?	bīs-la?bə-liA-m
ge	I.NOM	foot-GEN.1SG	underpart-LAT/LOC.3SG	water-PL	drink-DUR-PRS-1SG
mc	pers	n-n:case-poss	n-n:case-poss	n-n:num	v-v>v:tense-v:pn
fe	"I'm drinking waters down at my feet."				

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

(10)

ref	AA_1914_Hare_flk.012 (001.012)				
tx	"Mǎn	üjünə	jilgəndə	büjə?	bītle?bəliem."
mb	mǎn	üjü-nə	jil-gəndə	bü-jə?	bīt-le?bə-lie-m
mp	mǎn	üjü-nə	il-gəndə	bü-jə?	bīs-la?bə-liA-m
ge	I.NOM	foot-GEN.1SG	underpart-LAT/LOC.3SG	water-PL	drink-DUR-PRS-1SG
mc	pers	n-n:case-poss	n-n:case-poss	n-n:num	v-v>v-v:tense-v:pn
ps	pers	n	n	n	v
fe	"I'm drinking waters down at my feet."				

(11)

ref	AA_1914_Maneater_flk.012				
tx	T'üpi	pai?	pajbi,	šügəndə	embi.
mb	t'üpi	pa-i?	pa-j-bi	šü-gəndə	em-bi
mp	t'üpi	pa-jə?	pa-j-bi	šü-gəndə	hen-bi
ge	moist.[NOM.SG]	wood-PL	wood-VBLZ-PST.[3SG]	fire-LAT/LOC.3SG	put-PST.[3SG]
mc	adj.[n:case]	n-n:num	n-n>v-v:tense.[v:pn]	n-n:case-poss	v-v:tense.[v:pn]
ps	adj	n	v	n	v
fe	She chopped raw wood and put it onto her fire.				

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

Table 2: Tags for referent expressions

Tag	Description
0.1	zero/covert first-person referent
0.2	zero/covert second-person referent
0.3	zero/covert third-person referent

adv	adverbial referent
np	nominal referent (noun phrase)
pp	postpositional phrase
pro	pronominal referent
.h	human referent

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

Table 3: Semantic Roles tagged and their abbreviations

Semantic Role	Abbreviation	Description
Agent	A	- volitional initiator of the action - the participant which is volitionally causing the action - can be both animate and inanimate - test agent vs. theme: add "on purpose" to the sentence - if it fits, then it is an agent, if not, then not
Beneficiary	B	- entity for whose benefit the action is performed
Cause	Cau	- entity (mostly non-human) that causes an event
Comitative	Com	- entity that conveys a participant of the action (a.k.a. co-agent)
Experiencer	E	- entity that experiences the action or event - does not have a control over the action or event - verba sentiendi, i.e. verbs expressing emotion, volition, cognition, perception (i.e. verbs like: <i>see, love, hate, understand, hear, taste, frighten, wish, want, think, remember, feel</i>)
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 - arguments of unaccusative verbs such as <i>die, fall</i>
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 <i>see, hear, feel</i> , but NOT of verbs like <i>look for, listen</i>

Semantic Role	Abbreviation	Description
Theme	Theme	- 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 (usually) 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	- time point or an interval of time

The following charts shows some examples of tagging Semantic Roles:

(12)

ref	AA_1914_Maneater_flk.010 (001.010)		
tx	Mījebə	edəlabəʔbi	ked'i.
mb	mīje-bə	edə-labəʔ-bi	ked'i
mp	mīje-bə	edə-labəʔ-bi	ked'i
ge	soup-ACC.3SG	hang.up-RES-PST.[3SG]	away
ps	n	v	adv
SeR	np:Th 0.3.h:Poss	0.3.h:A	
fe	She hanged her soup away.		

(13)

ref	AA_1914_Maneater_flk.020 (001.020)		AA_1914_Maneater_flk.021 (001.021)	
tx	Dī	ne	nörbəlie:	"Püʔbdi!"
mb	dī	ne	nörbə-lie	püʔbdi-t
mp	dī	ne	nörbə-liA	püʔbdə-t
ge	this.[NOM.SG]	woman.[NOM.SG]	tell-PRS.[3SG]	blow-IMP.2SG.O
ps	dempro	n	v	v
SeR		np.h:A		0.2.h:A 0.3:P
fe	The woman says:		"Blow it!"	

2.10.3.8. Syntactic Functions (SyF)

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

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

verbal predicates the cells of the main verb and the auxiliary are merged. The following tags are used:

Table 4: Tags for annotating syntactic functions

Abbreviation	Description
Subject	
pro.h:S	pronominal human subject
pro:S	pronominal non-human subject
np.h:S	nominal human subject
np:S	nominal non-human subject
0.1.h:S	zero/covert first-person human subject
0.2.h:S	zero/covert second-person human subject
0.3.h:S	zero/covert third-person human subject
0.3:S	zero/covert third-person non-human subject
Direct Object	
pro.h:O	pronominal human direct object
pro:O	pronominal non-human direct object
np.h:O	nominal human direct object
np:O	nominal non-human direct object
0.3.h:O	zero/covert third-person human object
0.3:O	zero/covert third-person non-human object
Predicate	
v:pred	verbal predicate
n:pred	nominal predicate
adj:pred	attributive/adjectival predicate
pro:pred	pronominal predicate
ptcl:pred	particle predicate
cop	copula

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

Table 5: Tags for annotating subordinate clauses

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

The following charts show some examples of tagging syntactic functions:

(14)

ref	AA_1914_Maneater_flk.010 (001.010)		
tx	Mijebə	edəlabə?bi	ked'i.
mb	mijə-bə	edə-labə?-bi	ked'i
mp	mijə-bə	edə-labə?-bi	ked'i

ge	soup-ACC.3SG	hang.up-RES-PST.[3SG]	away
ps	n	v	adv
SyF	np:O	v:pred 0.3.h:S	
fe	She hanged her soup away.		

(15)

ref	AA_1914_Khan_flk.004 (001.004)		
tx	Aksa?	ibi	sejmüt.
mb	aksa?	i-bi	sejmü-t
mp	aksa?	i-bi	sejmü-t
ge	lame.[NOM.SG]	be-PST.[3SG]	mare-NOM/GEN.3SG
ps	adj	v	n
SyF	adj:pred	cop	np:S
fe	Lame was his mare.		

(16)

ref	AA_1914_Khan_flk.025 (002.003)				
tx	[...] kük	no?bə	tonəlla?bə	ej	surarga."
mb	kük	no?-bə	tonə-l-la?bə	ej	surar-ga
mp	kük	no?-bə	tonə-l-la?bə	ej	surar-ga
ge	green.[NOM.SG]	grass-ACC.3SG	tread-FRQ-DUR.[3SG]	NEG	ask-PTCP.PRS
ps	adj	n	v	ptcl	v
SyF		np:P	0.3.h:S v:pred	s:adv	
fe	[He hobbled his horse,] steps on my green grass without asking				

2.10.3.9. Information status (IST)

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

Table 6: Parameters for determining information status

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

In detail that means that given referents are necessarily and per default aforementioned in the discourse while accessible and new referents are not. Accessible referents can be somehow (see

below) inferred by the “hearer” of the discourse. Hence, new referents are neither aforementioned nor inferable for the hearer. The basic tags for annotating information status are *giv*, *accs* and *new*, the extended tag set can be seen from the following table:

Table 7: Basic tags for annotating information status

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

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

Another problem which was dealt with is the issue of direct speech. As is widely known, direct speech tends to change to perspective of both the hearer and the speaker which has consequences for the discourse status of referents as well. Simply spoken, a referent in direct speech has got an information status within the whole discourse/communication (i.e. for the hearer of the whole communication) and an information status within the micro-discourse made up with the usage of direct speech (i.e. for the hearer of the direct speech). As fine-grade discourse analysis is not the main goal of the project and would be very time-consuming, we decided to tag the information status of referents in direct speech on the level of the macro-discourse, i.e. the whole communication. However, in order to be aware of possible changes of perspective the tag <-Q> is added, if a referent occurs in direct speech (e.g. accs-gen-Q, i.e. a referent, accessible through general knowledge in direct speech), as it is done in the Nganasan Spoken Language Corpus (NSLC) [Brykina et al. 2018] according to Wagner-Nagy et al. [2018: 30]. Furthermore, so-called utterance predicates are tagged by the tag *quot* and it is distinguished between speech and thought (*quot-sp* vs. *quot-th*). The following examples show how information status is tagged:

(17)

ref	AA_1914_Khan_flk.003 (001.003)		
tx	Bile	sejmübə	d’appi.
mb	bile	sejmü-bə	d’ap-pi
mp	bile	sejmü-bə	d’abə-bi
ge	bad.[NOM.SG]	mare-ACC	capture-PST.[3SG]
ps	adj	n	v

IST		new	0.giv-active
fe	He caught a bad mare [for himself].		

(18)

ref	AA_1914_Khan_flk.004 (001.004)		
tx	Aksa?	ibi	sejmüt.
mb	aksa?	i-bi	sejmü-t
mp	aksa?	i-bi	sejmü-t
ge	lame.[NOM.SG]	be-PST.[3SG]	mare-NOM/GEN.3SG
ps	adj	v	n
IST			giv-active
fe	Lame was his mare.		

(19)

ref	AA_1914_Khan_flk.054 (002.032)			
tx	Ka:n	kegəriere:	“T’art’a bart’a	pol bart’a,
mb	ka:n	kegərer-ie	T’art’a bart’a	pol bart’a
mp	ka:n	kegərer-liA	tard’žabərd’ža	pol bart’a
ge	khan.[NOM.SG]	call.out-PRS.[3SG]	Tardzha-Bardzha	Pol-Bardzha
ps	n	v	propr	propr
IST	giv-inactive	quot-sp	giv-inactive-Q	giv-inactive-Q
fe	The khan shouts: “Tardzha-Bardzha Pol-Bardzha [...]”			

2.10.3.10. Borrowing (BOR)

The Borrowing tier (BOR) contains the annotation of borrowed lexical items. Both the origin of the item in question and the type of borrowing is annotated. The tags are made up as follows: <LANGUAGE:type>. The annotation is implemented already in the FLEx lexicon and automatically exported to EXMARaLDA. For Kamas, Russian (RUS) and Turkic (TURK) borrowings are distinguished, as well as a few occasional Mongolic (MONG), Khakas (KHAK) and Estonian (EST) ones. For the type of borrowing the following tags are used (cf. also Arkhipov 2020):

Table 8: Tags for annotating borrowings

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

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

2.10.3.11. Borrowing phonology (BOR-Phon)

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

Table 9: Annotation panel for phonological processes in borrowings

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

2.10.3.12. Borrowing morphology (BOR-morph)

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

Table 10: Tags for annotating morphological processes in borrowings

Tag	Description
Adaptation strategies	
dir:	direct insertion (i.e. insertion without morphological adaptation)
indir:	indirect insertion (i.e. insertion with morphological adaptation)
parad:	paradigm insertion (i.e. an inflected paradigm item is borrowed)
Further inflection (in the matrix language)	
:bare	no inflection
:infl	further inflection

An example of annotations in BOR tiers follows:

(20)

ref	PKZ_1964_SU0205.194 (044.004)						
ts	Süt izittə nada, ipek izittə, tus izittə, munuj? izittə.						
tx	Süt	izittə	nada,	ipek	izittə,	tus	izittə, ...
mb	süt	i-zittə	nada	ipek	i-zittə	tus	i-zittə
ge	milk.[NOM.SG]	take-INF.LAT	need.to	bread.	take-INF.LAT	salt	take-INF.LAT

				[NOM.SG]			
BOR	TUR:cult		RUS:mod	TUR:cult		TUR:cult	
BOR-Morph	dir:bare		dir:bare	dir:bare		dir:bare	
fe	One should take milk, take bread, take salt, take eggs.						

In Russian borrowings marked for plural, the following cases are distinguished (cf. Arkhipov 2023):

Plural marking strategy	Example	BOR tag
• Russian Nom. sg. stem + Kamas PL	v'inog-əʔi 'wreathes'	dir:infl
• Russian Nom. sg. with stress shift + Kamas PL	č'ola-ʔi 'bees'	indir:infl
• Russian Nom. sg. or pl. stem + Kamas PL	a'en'(-)ə(-)ʔi 'reindeer (pl.)	dir:infl
• Russian Nom. pl. stem + Kamas PL	ka'l'osa-ʔi 'wheels'	parad:infl
• Russian Nom. pl. stem	salav'ji 'nightingales'	parad:bare
• Russian generic Oblique stem + Kamas PL	bat'inka-ʔi 'shoes'	indir: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 language of the code-switch and type of the code switch are annotated, namely according to the scheme <LANGUAGE:type>. The language is mostly Russian (RUS). The tag set for the type of code-switch is the following:

Table 11: Tags for annotating code-switching

Tag	Description
Sentence-external code-switching	
:ext	languages change at sentence (clause, utterance) borders
Sentence-internal code-switching	
:int.ins	languages change at phrase borders (e.g. an NP or a PP is inserted)
:int.alt	the point of change is somewhere at an arbitrary point in the sentence
:int	a single word is inserted, distinguishing between subtypes is problematic

Code-switching to Russian is quite frequent in Plotnikova's recordings:

(21)

ref	PKZ_1964_SU0206.077 (017.004)				PKZ_1964_SU0206.078 (017.005)				
tx	Dīzeŋ	bar	kula:m̩biʔi	bar.	Ну,	наверно	все	тапер	уже
mb	dī-zeŋ	bar	ku-la:m̩-bi-ʔi	bar		е		ь	.
ge	s/he-PL	PTCL	die-RES-PST-3PL	all					
CS					RUS:ext				
fe	But they all died.				That's probably all by now.				

ref	PKZ_196X_FireBird_flk.077 (079)			
tx	"Dăk	tăn	tojirzittə	zăxatel? "
mb	dăk	tăn	tojir-zittə	
ge	so	you.NO M	steal-INF.LAT	
CS				RUS:int
fe	"So you wanted to steal?"			

2.10.3.14. Existential, locative and possessive clauses (ExLocPoss)

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

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

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

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

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

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

Table 12: Tags for annotating existential, locative and possessive clauses

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

:Ex.Aff	affirmative existential item as linking element
:Ex.Neg	negative existential item as linking element
:PosV.Aff	posture verb as linking element; affirmative
:PosV.Neg	posture verb as linking element; negative

As shown in Table 18, “zero copula” includes instances with and without pn-agreement at the predicate, i.e. the ground or possessee element. “Copula” means any linking item that has bleached semantics, regardless of whether it is nominal or verbal; most often, this item occurs in nominal or adjectival predication, as well. In Kamas, these are the copulas *i-* ‘be’ and *mo-* ‘become’. For tagging a linking element as “existential item”, it must be clear from a language-internal perspective that the item indeed has existential semantics – as for Kamas, this holds only for the negative existential item *naga* ~ *nago-*. Posture verbs are translational equivalents of English *sit, stand, lie, hang* etc.; note that clauses including posture verbs are only annotated as existential or locative clauses, if the figure element is inanimate. This is due to the fact that animate figures most often imply a semantic description of the posture instead of merely predicating episodic presence/absence (see Däbritz 2023 for a discussion). *have*-verbs are transitive translational equivalents for English *have*, inceptive verbs are verbs formed from a root meaning ‘find’, and other verbs subsume mostly verbs like ‘be situated’ or alike; these three possibilities are not encountered in the corpus.

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

(22)

ref	AA_1914_Mouse_flk.026 (001.028)		
tx	"Mǎn	dǐgǎn	igem."
mb	mǎn	dǐgǎn	i-ge-m
ge	PRO1SG.NO M	here	be-PRS2-1SG
ExLocPo ss	Loc:Cop.Aff		
fe	I am here.		

(23)

ref	PKZ_196X_SU0229.176 (176)		
tx	Magazingəndə	ĩmbi	naga, [...].
mb	magazin-gəndə	ĩmbi	naga
ge	shop-LAT/LOC.3SG	what. [NOM.SG]	NEG.EX.[3SG]
ExLocPo ss	Ex:Ex.Neg		
fe	There is nothing in the shop.		

(24)

ref	PKZ_196X_DaughterAndStepdaughter_flk.003 (003)				
tx	[...]	i	dǐ	büz'en	ko?bdo.
mb		i	dǐ	büz'e-n	ko?bdo
ge		and	this	man-GEN	girl.[NOM.SG]

ExLocP ss	Poss:Zero.Aff
fe	[...] and this man [had] a daughter.

2.10.3.15. Free translation (fe, fg, fr)

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

The following chart shows an example:

(25)

ref	AA_1914_Khan_flk.004 (001.004)		
tx	Aksa?	ibi	sejmüt.
mb	aksa?	i-bi	sejmü-t
ge	lame.[NOM.SG]	be-PST.[3SG]	mare-NOM/GEN.3SG
fe	Lame was his mare.		
fg	Seine Stute lahmte.		
fr	Хромой была его кобыла.		

2.10.3.16. Literal translation (ltg)

The literal translation tier (ltg) gives the original translation from [Joki 1944]. This translation represents an archaic variant of German and is mostly very close to the Kamas original. The following chart shows an example:

(26)

ref	AA_1914_Khan_flk.004 (001.004)		
tx	Aksa?	ibi	sejmüt.
mb	aksa?	i-bi	sejmü-t
mp	aksa?	i-bi	sejmü-t
ge	lame.[NOM.SG]	be-PST.[3SG]	mare-NOM/GEN.3SG
fe	Lame was his mare.		
fg	Seine Stute lahmte.		
ltg	<i>hinkend war sein Stute.</i>		
fr	Хромой была его кобыла.		

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

(27)

ref	PKZ_196X_SU0215.048 (051)	
tx	P'e	šobi.
mb	p'e	šo-bi
ge	year.[NOM.SG]	come-PST.[3SG]
fe	Summer came.	
nt	[GVY:] PKZ uses kö 'winter' as 'year' and p'e 'year' as 'summer'.	

References

- Arkhipov, Alexandre. 2023. Two morphologies, two stress systems, shaken, not stirred: Number marking on Russian borrowings in Kamas. In: D. Arbes (ed.) *Number Categories: Dynamics, Contact, Typology*. Berlin, Boston: De Gruyter Mouton. P. 63–106. <https://doi.org/10.1515/9783110986600-003>
- Arkhipov, Alexandre. 2020. INEL Corpora General Transcription and Annotation Principles. *Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology* 5. Szeged; Hamburg. DOI: [10.14232/wpcl.2020.5](https://doi.org/10.14232/wpcl.2020.5)
- Arkhipov A. V., Däbritz C. L. 2018. Hamburg corpora for indigenous Northern Eurasian languages // *Tomsk Journal of Linguistics and Anthropology*. 2018. Issue 3 (21). P. 9–18. URL: https://ling.tspu.edu.ru/en/archive.html?year=2018&issue=3&article_id=7130
- Brykina Maria, Gusev Valentin, Szeverényi Sandor, Wagner-Nagy Beáta. Nganasan Spoken Language Corpus (NSLC). Archived in Hamburger Zentrum für Sprachkorpora. Version 0.2. Publication date 12.06.2018. URL: <http://hdl.handle.net/11022/0000-0007-C6F2-8>
- Däbritz, Chris Lasse. 2023. Non-canonical structures in locative and existential predication in the Ob-Ugric languages. *Suomalais-Ugrilainen Seuran Aikakauskirja* 99. 244–274.
- 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: 20.12.2023].
- Haig, Geoffrey and Stefan Schnell 2014: *Annotations using GRAID (Grammatical relations and animacy in discourse)*, Introduction and guidelines for annotators, Version 7.0, Available online at https://multicast.aspra.uni-bamberg.de/data/pubs/graid/Haig+Schnell2014_GRAID-manual_v7.0.pdf [Accessed: 20.12.2023].
- Haspelmath, Martin. 2022. *Nonverbal clause constructions*. (Submitted manuscript.) (<https://ling.auf.net/lingbuzz/006673>) (Accessed 20.12.2023)
- Janurik, Tamás. 2014. *Kamassz szövegek Ago Künnap gyűjtéséből*. Székesfehérvár: Budenz alkotóház.
- Joki, Aulis Johannes. 1944. *Kai Donners Kamassisches Wörterbuch nebst Sprachproben und Hauptzügen der Grammatik*. Lexica Societatis Fenno-Ugricae 8. Helsinki: Suomalais-Ugrilainen Seura.
- Koch, Peter. 2012. Location, existence, and possession: A constructional-typological exploration. *Linguistics* 50(3). 533–603. (doi:[10.1515/ling-2012-0018](https://doi.org/10.1515/ling-2012-0018))
- Künnap, Ago. 1976a. Kamassilaisia tekstejä I. *Fenno-Ugristica* 2, 116–133. [KT I]
- Künnap, Ago. 1976b. Kamassilaisia tekstejä II. *Fenno-Ugristica* 3, 128–136. [KT II]
- Künnap, Ago. 1986. Kamassilaisia tekstejä III. *Fenno-Ugristica* 13, 166–173. [KT III]
- Künnap, Ago. 1990. Kamassilaisia tekstejä IV. *Fenno-Ugristica* 17, 218–237. [KT IV]
- Künnap, Ago. 1992a. Kamassilainen itkuvirsi 1914 ja 1965. *Fenno-Ugristica* 18, 119–122.
- Künnap, Ago. 1992b. Kamassilaisia arvoituksia. *Fenno-Ugristica* 18, 123–127.
- Künnap, Ago. 1992c. Kamassilaisia tekstejä V. *Fenno-Ugristica* 18, 128–140. [KT V]
- Wagner-Nagy, Beáta; Szeverényi, Sándor; Gusev, Valentin. 2018. User's Guide to Nganasan Spoken Language Corpus. // *Working Papers in Corpus Linguistics and Digital Technologies: Analyses*

and Methodology. Vol. 1, no. 1. P. 1-45. <https://doi.org/10.14232/wpcl.2018.1>. [Accessed: 20.12.2023].

Appendix 1. Kamas sounds and representing characters

Sound	Description	Unicode character names
a	low central unrounded vowel	latin small letter a
ă	short low central unrounded vowel	latin small letter a with breve
e	mid front unrounded vowel	latin small letter e
ə	mid central unrounded vowel	latin small letter schwa
i	high front unrounded vowel	latin small letter i
ĩ	short high front unrounded vowel	latin small letter I with breve
ɨ*	high central unrounded vowel	latin small letter i with stroke
o	mid back rounded vowel	latin small letter o
ö	mid front rounded vowel	latin small letter o with diaeresis
u	high back rounded vowel	latin small letter u
ü	high front rounded vowel	latin small letter u with diaeresis
b	voiced bilabial stop	latin small letter b
b'	palatalized voiced bilabial stop	latin small letter b + modifier letter apostrophe
p	voiceless bilabial stop	latin small letter p
p'	palatalized voiceless bilabial stop	latin small letter p + modifier letter apostrophe
f*	voiceless labial fricative (usu bilabial)	latin small letter f
v*	labial-velar glide [w] corresponding to Russian labiodental fricative [v]	latin small letter v
v'	palatalized voiced labial fricative	latin small letter v + modifier letter apostrophe
m	bilabial nasal	latin small letter m
m'	palatalized bilabial nasal	latin small letter m + modifier letter apostrophe
d	voiced dental stop	latin small letter d
d'	voiced palatal stop/affricate	latin small letter d + modifier letter apostrophe
t	voiceless coronal stop	latin small letter t
t'	voiceless palatal stop/affricate	latin small letter t + modifier letter apostrophe
s	voiceless alveolar fricative	latin small letter s
s'	voiceless palatal fricative	latin small letter s + modifier letter apostrophe
z	voiced alveolar fricative	latin small letter z
z'	voiced palatal fricative	latin small letter z + modifier letter apostrophe
č*	voiceless postalveolar affricate	latin small letter c with caron
š	voiceless postalveolar fricative	latin small letter s with caron
š'	palatalized voiceless postalveolar fricative	latin small letter s with caron + modifier letter apostrophe
ž	voiced postalveolar fricative	latin small letter z with caron
ž'	palatalized voiced postalveolar fricative	latin small letter z with caron + modifier letter apostrophe
n	alveolar nasal	latin small letter n
n'	palatal nasal	latin small letter n + modifier letter apostrophe
r	voiced dental trill	latin small letter r
l	dental lateral	latin small letter l
l'	palatalized dental lateral	latin small letter l + modifier letter apostrophe
j	palatal approximant	latin small letter j

g	voiced velar stop	latin small letter g
k	voiceless velar stop	latin small letter k
k'	palatalized voiceless velar stop	latin small letter k + modifier letter apostrophe
ŋ	velar nasal	latin small letter eng
h	voiceless glottal fricative	latin small letter h
h'	palatalized voiceless glottal fricative	latin small letter h + modifier letter apostrophe
ʔ	voiceless glottal stop	latin letter glottal stop

*Sounds only encountered in loanwords

: [modifier letter triangular colon] is used to mark long vowels.

Special transcription marks

(xyz)	Single brackets without trailing punctuation: uncertain transcription
(xyz-)	Single brackets with trailing dash: false starts rejected by the speaker, incomplete word
(xyz=)	Single brackets with trailing equals sign: false starts rejected by the speaker, complete word
(())	Double brackets contain special marks (see below):
((...))	Unclear fragment (not transcribed)
((BRK))	Break in recording (tape restarted)
((DMG))	Damaged recording (incl. on restarting the tape)
((COUGH))	Coughing
((LAUGH))	Laughing
((NOISE))	Noise
((PAUSE))	Long pause

For further details, please refer to the guidelines (Arkhipov 2020).

Appendix 2. Morpheme glossing labels (ge, gr)

Gloss	Value
1	first person
2	second person
3	third person
ABL	ablative
ACC	accusative
ADJZ	adjectivizer
CAR.ADJ	caritive adjective
CNG	connegative
COLL	collective
COM	comitative
COND	conditional converb
CVB	converb
CVB.ANT	converb of anteriority
CVB.NEG	negative converb
DAT	dative
DES	desiderative
DETR	detransitivizer
DIR	directive
DISC1	discourse particle <i>bar</i> (from Turkic)
DISC2	discourse particle <i>že</i> (from Russian)
DISC3	discourse particle <i>už</i> (from Russian)
DISC4	discourse particle <i>da</i> (from Russian)
DISC5	discourse particle = <i>da</i> , = <i>de</i> (from Turkic)
DISC6	discourse particle = <i>ta</i> , = <i>to</i> (from Russian - <i>mo</i>)
DISTR	distributive
DRV	(unspecified) derivation
DU	dual
DUR	durative
DUR.PRS	present durative
DUR.PST	past durative
DYA	dyadic
EP	epenthetic element
EX	existential
FACT	factive
FRQ	frequentative
FUT	future tense
GEN	genitive
HAB	habitual
HORT	hortative
IMP	imperative
INCH	inchoative
INDEF	indefinite

INF	infinitive
INS	instrumental
INS.N	instrumental nominal
INTJ	(unspecified) interjection
IPFVZ	imperfectivizer
IRREAL	irreal
JUSS	jussive
LAT	lative
LAT.ADV	adverbial lative
LOC	locative
LOC.ADV	adverbial locative
MOM	momentaneous
MULT	multiplicative
NEG	negative particle
NEG.AUX	negative auxiliary
NEG.EX	negative existential
NOM	nominative
O	objective conjugation
OPT	optative
ORD	ordinal number
PL	plural
POSS	possessive suffix unmarked for case
PREF	(unspecified loan) prefix
PRS1	present tense (-liA)
PRS2	present tense (-gA)
PST	past tense
PTCP	participle
PTCP.PASS	passive participle
PTCP.PRS	present participle
QP	question particle
RES	resultative
RFL	reflexive
SEM	semelfactive
SG	singular
TR	transitivizer
TR.RES	resultative transitivizer
VBLZ	verbalizer
%%	(unknown)

Appendix 3. Morphological category labels (ps, mc)

adj	Adjective
adv	Adverb
adv:case	Adverbial case marker
aux	Auxiliary
conj	Conjunction
dempro	Demonstrative pronoun
interj	Interjection
n	Noun
n:case	Nominal case marker
n:case.pos s	Nominal case marker, possessive declension
n:ins	Nominal epenthetic element
n:num	Nominal number marker
num	Numeral
pers	Personal pronoun
post	Postposition
prep	Preposition
propr	Proper Noun
ptcl	Particle
quant	Quantifier
que	Interrogative word
refl	Reflexive pronoun
v	Verb
v:ins	Verbal epenthetic element
v:mood	Verbal mood marker
v:mood.pn	Verbal mood marker specified for person-number
v:n.fin	Verbal non-finite marker
v:pn	Verbal (person-number) agreement marker
v:tense	Verbal tense marker
XX>YY	Derivational suffix which converts XX into YY (e.g. "n>v" is a verbalizer applied to nouns)
%%	(unknown)