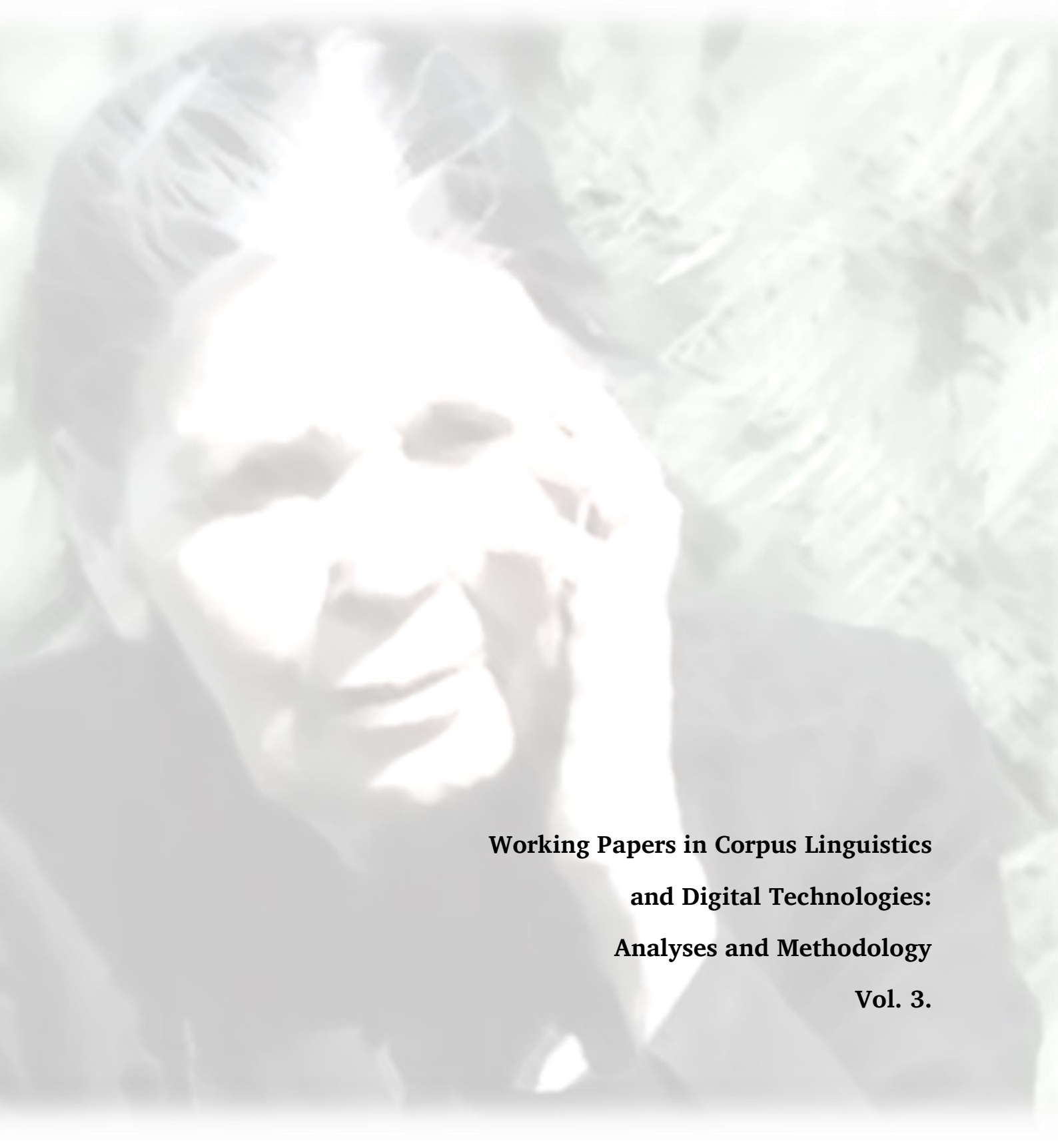
Alexandre Arkhipov – Chris Lasse Däbritz – Valentin Gusev

User's Guide to INEL Kamas Corpus



Working Papers in Corpus Linguistics
and Digital Technologies:
Analyses and Methodology
Vol. 3.

Alexandre Arkhipov – Chris Lasse Däbritz – Valentin Gusev

User's Guide to

INEL Kamas Corpus

Working Papers in Corpus Linguistics and Digital Technologies:

Analyses and Methodology

Vol. 3.

Szeged – Hamburg

2020

Working Papers in Corpus Linguistics and Digital Technologies: Analyses and methodology

Vol. 3

WPCL issues do not appear according to strict schedule.

Vol. 3 (2020)

Published by
University of Szeged, Department of Finno-Ugric Studies
Egyetem utca 2. 6722 Szeged
Universität Hamburg, Zentrum für Sprachkorpora
Max-Brauer-Allee 60 22765 Hamburg

Published 2020

ISBN 978-963-306-720-8 (pdf)

DOI: 10.14232/wpcl.2020.3

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.

1.2.2. 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.

1.2.3. Language Codes

ISO-639-3 code: **xas**
Glottolog code: **kama1378**

1.2.4. 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 data curation, archiving and publication are performed by the *Hamburg Centre for Language Corpora* (HZSK).[3]

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

[1] http://exmaralda.org/en/, last access: 18.12.2018

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

[3] https://corpora.uni-hamburg.de/hzsk/en, last access: 18.12.2018.

[4] https://creativecommons.org/licenses/by-nc-sa/4.0/, last access: 18.12.2018

1.4. Citation

The corpus is to be cited as follows:

> Gusev, Valentin; Klooster, Tiina; Wagner-Nagy, Beáta. 2019. INEL Kamas Corpus. Version 1.0. Publication date 2019-12-15. http://hdl.handle.net/11022/0000-0007-DA6E-9. Archived in Hamburger Zentrum für Sprachkorpora. In: Wagner-Nagy, Beáta; Arkhipov, Alexandre; Ferger, Anne; Jettka, Daniel; Lehmberg, Timm (eds.). The INEL corpora of indigenous Northern Eurasian languages.

1.5. Project members

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 research was carried out at the Institute for Finno-Ugric/Uralic Studies (IFUU) of the Universität Hamburg (UHH). The technical infrastructure was provided by the Hamburger Zentrum für Sprachkorpora (HZSK). The project homepage can be visited at: https://inel.corpora.uni-hamburg.de/. Contributions of particular researchers are acknowledged in more detail in the metadata to the corpus (see 2.1.3).

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.

Developers

Timm Lehmberg, M.A., Technical coordinator
Daniel Jettka, M.A.
Niko Partanen, M.A. (February 2016 – March 2017)
Anne Ferger, M.A. (since April 2017)

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)
Theodor Hey (April 2019 – September 2019)
Felicitas Otte (April 2019 – December 2019)

1.6. Acknowledgements

1.6.1. Funding

This corpus has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry

of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.[5]

1.6.2. Data sources

Texts collected by Kai Donner come from an unpublished edition of Kai Donner's manuscripts by 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. Some layers of information such as the German translation by A. Joki are taken from the published version of the texts (Joki 1944).

Gerson 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 Gerson 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 1.0

The totality of Klavdiya Plotnikova's transcripts are now published, including all the tapes from the KOTUS archive, as well as the two recordings of Aleksandra Semyonova (21 more texts in total).

All the texts are now annotated for syntactic functions and semantic roles.

Numerous corrections have been made in glosses and other annotations.

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.

[5] The project was applied for by Prof. Dr. Beáta Wagner-Nagy, Dr. Michael Rießler, Hanna Hedeland, M.A., and Timm Lehmberg, M.A.

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 Aulis 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 1.6.2), 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).

2.5. Corpus size

This version of the corpus contains 158 transcripts (102 folklore, 1 song, 11 narratives, 44 miscellaneous) of 4 known speakers (for 2 texts the speaker is unknown) with 13 790 sentences and 63 824 words including Russian code-switching and speech disfluencies. With Russian fragments and false starts excluded, the word count is approximately 45 000. 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 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. Long vowels are marked by the character **:** (triangular colon) following the vowel sound (e.g. *ša:škən* 'magpie'). Since the existence of inherent vowel length distinctions in Kamas is yet unclear, currently the only words where the long vowels are marked are words where long vowels have emerged as a result of contraction inside the stem. Long vowels are marked with double characters in the positions 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'ăga* '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 publication of Joki (1944), in PDF (*.pdf) (for texts from Donner's collection)
- sound file in WAV (*.wav) (for texts with audio source)

Supplementary folders:

- documentation (user documentation)

corpus-utilities (conversion settings and annotation panels used with EXB transcriptions)

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

2.6.6.3. 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:

2.6.6.4. Data collectors and editors

DK: Donner, Kai
JA: Joki, Aulis
KA: Künnap, Ago
KlG: Klumpp, Gerson
MAK: Matveev, Aleksandr
VTR: Viitso, Tiit-Rein
YPJ: Yli-Paavola, Jaakko

2.6.6.5. Project members

AAV: Arkhipov, Alexandre
BJ: Budzisch, Josefina
DCh: Däbritz, Chris Lasse
GVY: Gusev, Valentin
KlT: Klooster, Tiina
WH: Wegener, Hannah
WNB: Wagner-Nagy, Beáta

2.6.6.6. Student assistants

DO: Degtyareva, Olesya
JK: Jessen, Kis
KH: Klitzing, Hannes
OF: Otte, Felicitas

2.6.6.7. 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".

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 Jokis book (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)[6] for glossing.

For the post-shift collection, the sound files in wav format were segmented into utterances and transcribed in ELAN multimedia annotator[7], translated into English and then imported into FLEx for glossing.

For all transcripts, the morphological analysis (interlinear glossing) was done in FLEx. 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 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 (see also 2.10).

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 Joki.

---

[6] https://software.sil.org/fieldworks/, last access: 26.10.2017.

[7] https://tla.mpi.nl/tools/tla-tools/elan/, last access: 28.11.2017.

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)
- **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 2.6.6.3)
- **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 *ńigə* (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

The transcripts in EXB format contain multiple tiers (layers of information). The purpose of different tiers is outlined in Table 1 below and explained in more detail in subsequent sections. For texts which only have one speaker transcribed, the tier names are identical to the tier category as given in Table 1. For texts with more than one speaker, the actual tier names additionally include the speaker code separated by a dash (e.g. "ref-PKZ"). Each speaker thus can have a full set of tiers. In practice, however, there are no dialogues between Kamas speakers in the corpus, and all but one "target" speaker are researchers (or technical assistants) who speak other languages than Kamas (except for several utterances by Ago Künnap). They are thus annotated using only a subset of tiers, mostly sentence-level ones, namely: ref, ts, tx, CS, fr, fe, fg, nt.

Table 1 *Overview of tiers*

Tier category	*Tier full* ***name***	***Description***	***Unit***	***Optionality***
ref	Reference	Text ID + sentence number	sentence	obligatory
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

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

2.10.2. Transcription tiers (tx, ts)

The main transcription tiers (tx and ts) use the INEL Kamas transcription (see 2.6.6.2). 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 to?bdobi, nugurbinə püjebə băppi.				
tx	Kozan	kandəbi,	kandəbi.	Nugurbi	to?bdobi,	nugurbinə	püjebə	băppi.
fe	A hare walked and walked.			He met steppe grass, on the steppe grass he cut his nose.				

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

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; see chart (2). The sentences are numbered throughout the entire text, separately for each speaker. In brackets, the numbering according to the FLEx scheme is given (paragraph number followed by sentence number, irrespective of the speaker). The sentence and paragraph numbers are zero-padded up to 3 digits. For texts with more than one speaker, the speaker code is added after the communication code, also separated by a dot, as in chart **Hiba! A hivatkozási forrás nem található.**.

(2)

ref	AA_1914_Corpse_flk.073 (003.038)

(3)

ref	PKZ_1964_SU0207.KA.005 (012.004)

2.10.3.2. Morpheme breaks and morphophonemes (mb, mp)

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. Clitics are separated by equals sign. 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.

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. Zero morphs are not represented in this tier. The separators (dashes and equals signs) are identical in **mb** and **mp** tiers. For an example see chart **Hiba! A hivatkozási forrás nem található.** below.

(4)

ref	AA_1914_Hare_flk.012 (001.012)				
tx	"Măn	üjünə	jilgəndə	büjəʔ	bĭtleʔbəliem."
mb	măn	üjü-nə	jil-gəndə	bü-jəʔ	bĭt-leʔbə-lie-m
mp	măn	üjü-nə	il-gəndə	bü-jəʔ	bĭs-laʔbə-liA-m
fe	"I'm drinking waters down at my feet."				

2.10.3.3. Gloss (ge, gr)

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

Glosses for all morphs within a word are separated with hyphens. Glosses for zero morphs are given in square brackets preceded by a dot (e.g. go-PST.[3SG]). Glosses of clitics are separated by equals

[8] https://www.eva.mpg.de/lingua/resources/glossing-rules.php, last access: 01.11.2017.

signs (e.g. *ĭmbi=n'ibud'* what=INDEF). Some epenthetic elements in verbal and nominal inflection, while not being separate morphs, are segmented in **mb** and **mp** tiers and glossed as EP.

If a morpheme contains two or more semantic components, these are separated by a dot; the dot is omitted in combinations of person and number (e.g. IMP.2SG). Some inflected wordforms are left unsegmented, using the dot to separate the grammatical gloss, e.g. *măna* I.LAT (ge) / я.LAT (gr). The 2 pers. plural pronoun gets a combined gloss in English to distinguish it from the singular, cf.: *tănan* you.ACC (ge) / ты.ACC (gr), *ši?n'ile?* you.PL.ACC (ge) / вы.ACC (gr). Russian loanwords which are borrowed in an inflected form also get a combined gloss, see (7).

The dot is also used to join words in a multi-word lexical gloss, e.g. throw.away. Orthographical hyphen is replaced with an underscore, e.g. sister_in_law or Tardzha_Bardzha.

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

Morphemes with unknown meaning are glossed with two percent signs (%%), as in **Hiba! A hivatkozási forrás nem található.**. One leading percent sign indicates that the gloss is tentative (e.g. *kormiš* %spindle).

(5)

ref	AA_1914_Hare_flk.012 (001.012)				
tx	"Măn	üjünə	jilgəndə	büjə?	bĭtle?bəliem."
mb	măn	üjü-nə	jil-gəndə	bü-jə?	bĭt-le?bə-lie-m
mp	măn	üjü-nə	il-gəndə	bü-jə?	bĭs-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."				

(6)

ref	PKZ_196X_SU0228.115 (117)				
tx	Ma:?ndə	šobiam,	dĭ	inem	kolerbiem.
mb	ma:?-ndə	šo-bia-m	dĭ	ine-m	koler-bie-m
mp	ma?-gəndə	šo-bi-m	dĭ	ine-m	koler-bi-m
ge	house-LAT/LOC.3SG	come-PST-1SG	this.[NOM.SG]	horse-ACC	%%-PST-1SG
gr	дом-LAT/LOC.3SG	прийти-PST-1SG	этот.[NOM.SG]	лошадь-ACC	%%-PST-1SG
fe	I came home, (unharnessed?) the horse.				

(7)

ref	PKZ_196X_SU0218.040 (040)		
tx	Ăgărottə	de?	ăgurci?i.
mb	ăgărot-tə	de-?	ăgurci-?i
mp	ăgărot-Tə	det-?	ăgurci-jə?
ge	plant.garden-LAT	bring-IMP-1SG	cucumber.PL-PL
gr	огород-LAT	принести-IMP-1SG	огурец.PL-PL
fe	Bring the cucumbers [from] the vegetable garden.		

2.10.3.4. Morphological category (mc) and part of speech (ps)

The **ps** tier contains part of speech labels for each word form. For derivations, only the resulting part of speech is reflected in this tier. The **mc** tier indicates the morphological category for each morpheme, including zero morphs, joined with the same separators as in the glossing tiers (i.e. dash, dot with square brackets or equals sign). Both tiers are exported from FLEx along with the glosses.

For lexical stems, this is identical to the part of speech, using the same set of tags as in the **ps** tier.

For derivational affixes, the source and target part-of-speech labels are joined with a > sign, e.g. v>n for deverbal nominal derivation.

For inflectional affixes, the category is labeled with a host part-of-speech tag followed by an indication of categories expressed, separated with a colon, e.g. n:num for a nominal number marker; n:case.poss for a nominal possessive case marker, v:tense for a verbal tense marker. Please note that these category labels are very coarse-grained and do not always reflect the details of the morphological structure. For the list of abbreviations, see Appendix 3. The following chart shows an example of both tiers:

(8)

ref	AA_1914_Hare_flk.012 (001.012)				
tx	"Măn	üjünə	jilgəndə	büjəʔ	bĭtleʔbəliem."
mb	măn	üjü-nə	jil-gəndə	bü-jəʔ	bĭt-leʔbə-lie-m
mp	măn	üjü-nə	il-gəndə	bü-jəʔ	bĭs-laʔbə-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."				

(9)

ref	AA_1914_Maneater_flk.012				
tx	T'üpi	paiʔ	pajbi,	šügəndə	embi.
mb	t'üpi	pa-iʔ	pa-j-bi	šü-gəndə	em-bi
mp	t'üpi	pa-jəʔ	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.				

2.10.3.5. Semantic roles (SeR)

The Semantic roles tier (SeR) contains the annotation of semantic roles (a.k.a. thematic roles, theta-roles). The annotation is based on GRAID principles (Haig & Schnell 2014) with further developments by Beáta Wagner-Nagy and Sándor Szeverényi (Wagner-Nagy et al. 2018: 21ff.), adapted for the current project. The annotation takes into account form, animacy and semantic role of the referent, the tags are built up according to the scheme <form.animacy:semantic_role>. If the referent is expressed by a whole phrase, then the semantic role is tagged at the head of the phrase. In postpositional constructions the cells of the postposition and its complement are merged. 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 suffix <.h>, non-human referents get no additional marking. Table 2 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 distinguished semantic roles are explained in the following table.

Table 3 *Semantic roles and their abbreviations*

Semantic Role	**Abbr.**	**Description**
Agent	A	volitional initiator of the action; the participant which is volitionally causing the action can be both animate and inanimate; test agent vs. theme: add "on purpose" to the sentence – if it fits, then it is an agent, if not, then not
Beneficiary	B	entity for whose benefit the action is performed
Cause	Cau	entity (mostly non-human) that causes an event
Comitative	Com	entity that convoys a participant of the action (a.k.a. co-agent)
Experiencer	E	entity that experiences the action or event; does not have a control over the action or event; verba sentiendi, i.e. verbs expressing emotion, volition, cognition, perception (i.e. verbs like: *see, love, hate, understand, hear, taste, frighten, wish, want, think, remember, feel*)
Goal	G	location or entity in the direction of which something moves (i.e. directional location)
Instrument	Ins	medium by which the action or event is performed
Location	L	location or entity where an event takes or place or where something is located (i.e. stative location)
Path	Path	entity or location along or through which the event takes place
Patient	P	undergoer of the action; test patient vs. theme: does the referent change its quality during the action? – if yes, then patient; arguments of unaccusative verbs such as *die, fall*

Semantic Role	Abbr.	Description
Possessor	Poss	entity which owns something; both alienable and in-alienable possession; also inanimate referents (e.g. the top of the mountain)
Recipient	R	(mostly animate) recipient of transfer of something; addressee of verba dicendi
Source	So	location or entity where a movement starts (i.e. directional location); original owner in a transfer of something
Stimulus	St	stimulus for physical perception, i.e. second actant of verbs like *see*, *hear*, *feel*, but NOT of verbs like *look for*, *listen*
Theme	Theme	entity which is moved or affected by some action (change of location or possession; object of transfer); entity whose location is specified; 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

For examples of semantic roles annotations, see charts **Hiba! A hivatkozási forrás nem található.**–**Hiba! A hivatkozási forrás nem található.** in the next section.

2.10.3.6. Syntactic functions (SyF)

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

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

Table 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

Abbreviation	Description
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 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 and semantic roles:

(10)

ref	AA_1914_Maneater_flk.010 (001.010)		
tx	Mĭjebə	edəlaba?bi	ked'i.
mb	mĭje-bə	edə-laba?-bi	ked'i
mp	mĭje-bə	edə-laba?-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	
SyF	np:O	v:pred 0.3.h:S	
fe	She hanged her soup away.		

(11)

ref	AA_1914_Maneater_flk.020 (001.020)			
tx	Dĭ	ne	nörbəlie:	"Pü?bdit!"
mb	dĭ	ne	nörbə-lie	pü?bdi-t
mp	dĭ	ne	nörbə-liA	pü?bdə-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:Th
SyF		np.h:S	v:pred	v:pred 0.2:S 0.3:O
fe	The woman says:			“Blow it!”

(12)

ref	AA_1914_Khan_flk.024 (002.003)					
tx	[...] măn	kük	no?bə	tonəlla?bə	ej	surarga."
mb	măn	kük	no?-bə	tonə-l-la?bə	ej	surar-ga
mp	măn	kük	no?-bə	tonə-l-la?bə	ej	surar-ga
ge	I.GEN	green.[NOM.SG]	grass-ACC.3SG	tread-FRQ-DUR.[3SG]	NEG	ask-PTCP.PRS
ps	pers	adj	n	v	ptcl	v
SeR	pro.h:Poss		np:P	0.3.h:A		0.3.h:A
SyF			np:O	v:pred 0.3.h:S	s:adv	
fe	[He hobbled his horse,] steps on my green grass without asking.					

2.10.3.7. Information status (IST)

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

Table 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 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 Kamas is a pro-drop language, many referents are not overtly realized in the sentence. Therefore the information status of non-overt referents is also 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 suffix <-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:

(13)

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].		

(14)

ref	AA_1914_Khan_flk.004 (001.004)		
tx	Aksaʔ	ibi	sejmüt.
mb	aksaʔ	i-bi	sejmü-t
mp	aksaʔ	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.		

(15)

ref	AA_1914_Khan_flk.054 (002.032)			
tx	Kaːn	kegərerieː	“T’art’a bart’a	pol bart’a,
mb	kaːn	kegərer-ie	T’art’a bart’a	pol bart’a
mp	kaːn	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 […]”			

2.10.3.8. Borrowing (BOR)

The Borrowing tier (BOR) contains the annotation of borrowed lexical items. Both the source language of borrowing 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 Kamas, the source languages of annotated borrowings include Russian (RUS), unspecified Turkic (TURK) and, in some cases, Tatar (TAT). For the type of borrowing the following tags are used (cf. also Arkhipov 2020).

Table 8 *Tags for borrowing types*

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.

Most identified borrowings are lexical items, however there are also some borrowed bound morphemes, notably indefinite pronoun-forming markers and modal clitics. Since they normally appear in transcriptions attached to another lexical word, their annotation in the BOR tier additionally includes the corresponding gloss in brackets in order to distinguish it from the BOR annotation of lexical items which might occur in the same word, as in (16).

(16)

ref	SAE_196X_SU0231.SAE.151 (155)	
tx	Alləm	gibər-n'ibud'.
mb	al-lə-m	gibər = n'ibud'
ge	go-FUT-1SG	where.to = INDEF
BOR		RUS:gram(INDEF)
fe	I'll go somewhere.	

2.10.3.9. Borrowing phonology (BOR-Phon)

The tier *BOR-Phon* contains the annotation of phonological processes in borrowing. The tag set is shown in table 9.

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)

2.10.3.10. 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:

(17)

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	one.should	bread.[NOM.SG]	take-INF.LAT	salt	take-INF.LAT
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.						

2.10.3.11. 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 languages encountered in Kamas corpus are Russian (RUS), Estonian (EST), Finnish (FIN) and Khakas (KHAK). 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:

(18)

ref	PKZ_1964_SU0206.077 (017.004)				PKZ_1964_SU0206.078 (017.005)				
tx	Dĭzeŋ	bar	kula:mbiʔi	bar.	Ну,	наверное	все	теперь	уже.
mb	dĭ-zeŋ	bar	ku-la:m-bi-ʔi	bar					
ge	s/he-PL	PTCL	die-RES-PST-3PL	all					
CS					RUS:ext				
fe	But they all died.				That's probably all by now.				

(19)

ref	PKZ_196X_FireBird_flk.077 (079)			
tx	"Dăk	tăn	tojirzittə	zăxatel?"
mb	dăk	tăn	tojir-zittə	
ge	so	you.NOM	steal-INF.LAT	
CS				RUS:int
fe	So you wanted to steal?			

2.10.3.12. Free translation (fe, fg, fr) and literal translation (ltg)

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 a separate document (Arkhipov 2020).

The literal translation tier (**ltg**) gives the original translation from (Joki 1944). This translation represents an archaic variant of German and in some respects very close to the Kamas original. The following chart shows an example:

(20)

ref	AA_1914_Khan_flk.004 (001.004)		
tx	Aksaʔ	ibi	sejmüt.
mb	aksaʔ	i-bi	sejmü-t
mp	aksaʔ	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	*hinkend war sein Stute.*		
fr	Хромой была его кобыла.		

2.10.3.13. 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).

(21)

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 2020: INEL Corpora General Transcription and Annotation Guidelines. *Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology*.

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

Brykina, Maria & Valentin Gusev & Sándor Szeverényi & Beáta Wagner-Nagy 2018: *Nganasan Spoken Language Corpus (NSLC)*. Archived in Hamburger Zentrum für Sprachkorpora. Version 0.2. Publication date 12.06.2018. [Available online at http://hdl.handle.net/11022/0000-0007-C6F2-8]

Götze, Michael et al. 2007: Information structure, in Dipper, Stefanie & M. Götze & S. Skopeteas (eds): *Information Structure in Cross-Linguistic Corpora.* Interdisciplinary Studies on Information Structure 07: 147–187, [Available online at https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/deliver/index/docId/2036/file/Kapitel6_07.pdf] [Accessed: 02.11.2017].

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://www.uni-bamberg.de/fileadmin/aspra/Publications/GRAID7.0_manual.pdf] [Accessed: 01.11.2017].

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.

Künnap, Ago. 1976a: Kamassilaisia tekstejä I. *Fenno-Ugristica* 2: 116–133.

Künnap, Ago. 1976b: Kamassilaisia tekstejä II. *Fenno-Ugristica* 3: 128–136.

Künnap, Ago. 1986: Kamassilaisia tekstejä III. *Fenno-Ugristica* 13: 166–173.

Künnap, Ago. 1990: Kamassilaisia tekstejä IV. *Fenno-Ugristica* 17: 218–237.

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.

Wagner-Nagy, Beáta & Sándor Szeverényi & Valentin Gusev 2018: User's Guide to Nganasan Spoken Language Corpus. *Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology* 1: 1–45. [Available online at https://ojs.bibl.u-szeged.hu/index.php/wpcl/article/view/10611. [Accessed: 16.01.2019].

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
ĭ	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

Sound	Description	Unicode character names
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 details, please refer to the guidelines on transcription and translation (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
CVB	converb
CVB.ANT	converb of anteriority
CVB.NEG	negative converb
CVB.TEMP	temporal converb
DAT	dative
DES	desiderative
DETR	detransitivizer
DIR	directive
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	factitive
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

Gloss	Value
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	obejctive conjugation
OPT	optative
ORD	ordinal number
PL	plural
PREF	(unspecified loan) prefix
PRS	present tense
PST	past tense
PTCL	(unspecified) particle
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.poss	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)