INEL Kalmyk corpus

User documentation

Vlada Baranova, 17.07.2025

1. Introduction

1.1. Objective of the corpus

The present corpus of Kalmyk has been created as part of the long-term research project INEL ("*Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages*")[1]. Its primary goal is to create digital and machine-searchable corpora of several indigenous Northern Eurasian Languages.

The INEL Kalmyk corpus contributes to the documentation of the indigenous languages of Northern Eurasia and makes possible further research on the language. The Kalmyk language is, in general, fairly well described, and there exists a small corpus of written material, mainly derived from newspapers.[2] An extended written corpus is currently being developed at the Kalmyk Research Centre RAS,[3] though it is not searchable online. The corpus presented here differs in that it is oriented toward spoken language and includes morphological annotations (with homonymy resolved) and translations. As such, it can be used by both Kalmyk language specialists and linguistic typologists.

1.2. Kalmyk language

1.2.1. Description

Kalmyk is a Central Mongolian variety (Janhunen 2006: 232). This group includes Standard Mongolian (Khalkha), Buryat, Khamnigan, Khorchin, Ordos, and Oirat/Kalmyk. Other Oirat varieties are spoken in Xinjiang and Inner Mongolia, China, in western Mongolia, and in the Issyk Kyl province, Kyrgyzstan. Nomads settled over a large territory, and in the early 17th century, some Oirats formed an exclave in the Volga region, which was far to the west compared to other Mongol groups.

There is a discussion on the status of different Oirat varieties. Although many variants of the Oirat dialects are mutually intelligible, the development of these variants in isolation from each other (and Kalmyk in isolation from all other Mongolic languages) has resulted in the formation of distinct variants of the Oirat language. Studies of the written Oirat language refer to all these variants as Oirat (Rákos 2015). However, the self-identification of speakers is equally important, and the inhabitants of Kalmykia refer to their language as *Kalmyk*. The Kalmyks now live in the Republic of Kalmykia in southern Russia.

Typologically, Kalmyk (Oirat) is an agglutinative language with SOV word order. The case system has nominative, genitive, dative-locative, accusative, instrumental, associative, comitative, ablative and directional case suffixes. In verbal morphology, Kalmyk also has auxiliaries and periphrastic constructions expressing various aspectual and modal categories.

The alphabet used for Oirat/Kalmyk before the Soviet era, known as Clear Script or *Todo bichig* in Oirat, was created in the 17th century by an Oirat Buddhist monk named Zaya Pandita. In the 1920s-30s, the vertical script was replaced by Cyrillic, later switched to Latin letters, and then back to a slightly different version of Cyrillic with additional characters for specific Kalmyk sounds.

Kalmyk is the second official language of the Republic of Kalmykia. According to its constitution, Kalmyk is a language of official communication (along with Russian). Despite this, Kalmyk is classified as an endangered language by UNESCO. The Soviet-era Russification policies had a particularly harsh impact on the Kalmyk people. In 1943, they

[1] https://www.slm.uni-hamburg.de/inel/, last access: 07.07.2025.
[2] http://web-corpora.net/KalmykCorpus/search/?interface_language=ru, last access: 07.07.2025.
[3] http://kalmcorpora.ru/, last access: 07.07.2025.

were deported to Siberia, where they remained in exile until 1956, stripped of many civil rights. The conditions during their exile were extremely hard, leading to the deaths of many Kalmyks and language shift.

1.2.2. Language codes

ISO 639-3 code: **xal**

Glottolog code: **kalm1243**

1.2.3. Dialectal subdivisions

There are three major dialects of Kalmyk, Derbet, Torgut, and Buzava, which only exhibit some slight differences in lexicon and pronunciation. The standard language is mostly based on the Derbet dialect. Most of the texts in this corpus are also in the Derbet dialect, except for the records from the village of Sarpa (Torgut).

1.1. Archiving

The INEL Kalmyk corpus consists of texts provided with source media files (whenever available, 48 from 55) and annotated transcripts in *EXMARaLDA*[4] transcript format.

Texts are provided with metadata descriptions in *EXMARaLDA* Coma format.

For the texts with available recordings a copy of *EXMARaLDA* transcripts in ELAN[5]. EAF format is also provided as an alternative for ELAN users. A copy of transcripts in ISO/TEI format is provided for use in compatible tools, in particular for the Tsakorpus online search platform.

The corpus is archived and published by the Research Data Repository of the University of Hamburg[6] under open-access conditions with Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0).[7]

The corpus is available for download in three packages of different size:

- The "standard" package includes sound files in WAV format.
- The "mp3" package includes sound files in MP3 format.
- The "lite" package does not include any sound.

Besides the downloadable packages, the corpus is accessible online through Tsakorpus,[8] an open-source search platform for linguistic corpora. The current version of the corpus can be accessed at https://inel.corpora.uni-hamburg.de/KalmykCorpus/search.

1.2. Citation

Baranova, Vlada. INEL Kalmyk Corpus. Version 1.0. Publication date 2025-07-17. https://hdl.handle.net/11022/0000-0007-FFB1-2. Archived at Universität Hamburg. In: *The INEL Corpora of Indigenous Northern Eurasian Languages*. https://hdl.handle.net/11022/0000-0007-F45A-1

1.3. Project members

Project leader

Beáta Wagner-Nagy

Editors

Alexandre Arkhipov

Main corpus authors

Vlada Baranova

Technical developers

Aleksandr Riaposov

Elena Lazarenko

[4] http://exmaralda.org/en/, last access: 07.07.2025.
[5] https://tla.mpi.nl/tools/tla-tools/elan/, last access: 07.07.2025.
[6] https://www.fdr.uni-hamburg.de/communities/inel, last access: 07.07.2025.
[7] https://creativecommons.org/licenses/by-nc-sa/4.0/, last access: 07.07.2025.
[8] https://github.com/timarkh/tsakorpus, last accessed: 07.07.2025.

1.4. Acknowledgements

Funding

The final version of the corpus has been produced in Oct. 2024 – Ferb. 2025 in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities. The project was applied for by Prof. Dr. Beáta Wagner-Nagy, Dr. Michael Rießler, Hanna Hedeland, M.A., and Timm Lehmberg, M.A.

In the framework of the short-term fellowship "Corpus of texts in spoken Kalmyk language" in Oct. 2024 – Ferb. 2025, the texts were prepared; however, of course, the preliminary work of recording, translating, transcribing, and analyzing the texts took much more time and required the efforts of many people.

A number of texts were recorded and/or transcribed in 2008–2018 within projects supported by the following grants at the Institute for Linguistic Studies RAS:

- Russian Foundation for Basic Research grant "Typology of the Mechanisms of Interaction between Russian and the Languages of Small-numbered peoples of Russia" (#17-2909097, 2018–2021, PI Evgeny Golovko);
- Russian Foundation for Humanities grant "Language change in lects with no writing tradition (Altaic, Paleoasiatic and Uralic languages)" (#13-04-00416, 2013–2015, PI Evgeny Golovko);
- Russian Foundation for Basic Research grant "Creating corpora of texts in the languages of Russia: Nanai, Udihe, Kalmyk" (#07-06-00278a, 2007-2009, PI Elena Perekhvalskaya).

Text producing, text recording and analysis, processing the data

First, it is important to express deep gratitude to all native speakers who generously shared their knowledge of Kalmyk, making the creation of this corpus possible. Zamira Xejchieva and Galina Cabdy`rova assisted with oral transcription and the Russian translation of the audio materials.

I would like to express my special thanks to Sergey Say for his contribution to the development of the texts, transcriptions, and glosses in 2007–2009.

Part of the materials were recorded during joint expeditions of St. Petersburg University and the Institute for Linguistic Studies of the Russian Academy of Sciences in 2007–2008, under the direction of Elena Perekhvalskaya and Sergey Say. Deep gratitude to Sergey Say, Elena Perekhvalskaya, Ksenia Goto, Maria Konoshenko, Maria Ovsjannikova, Sofia Oskol`skaya, Ksenia Shagal and Aleksandra Vydrina† who participated in collecting or transcribing the texts in 2007-2008.

2. The corpus

2.1. Content

The corpus contains texts of various genres, which are broadly classified (following the INEL project conventions) as folklore, narrative (monologues that are not folklore texts) and conversation.

It includes texts consisting of a string of transcriptions of the Kalmyk text, glossing and translations into Russian and English. The texts are accompanied by audio files, except for a few cases where the audio has been lost.

2.2. Sources

Half of the materials were recorded during joint expeditions of St. Petersburg University and the Institute for Linguistic Studies of the Russian Academy of Sciences between 2006 and 2008, under the direction of Elena Perekhvalskaya and Sergey Say.

These texts were recorded in the villages of Ergeninskij and Tugtun in the Ketchenerovsky District (Derbet dialect) by Vlada Baranova, Ksenia Goto, Maria Konoshenko, Maria Ovsjannikova, Sofia Oskol`skaya, Elena Perekhvalskaya and Aleksandra Vydrina†. In 2008–2009, they were transcribed and analyzed by Sergey Say and Vlada Baranova, with several texts also processed by Elena Perekhvalskaya, Aleksandra Vydrina†, Ksenia Goto, and Ksenia Shagal. Zamira Xejchieva came to St. Petersburg in 2008 specifically to assist with the transcription work.

The transcription and glossing system were developed by the group of researchers led by Sergey Say with input from other participants. This system was used in the publication of texts (5,750 words, based on approximately 60 minutes of audio) (Baranova, Say 2009; Texts 2009). Editing and preparation of the materials were carried out by Say and Baranova with the help of other contributors, particularly Mariya Xolodilova.

This transcription system with some modifications in transcription and more substantial changes in glossing has also been used in the present corpus.

Texts from 2014, 2015, and 2018 were recorded by Vlada Baranova in the villages of Tugtun and E`vdy`k in the Ketchenerovskij district (Derbet dialect), and in the village of Sarpa (Torgut dialect). Texts were transcribed and translated by Baranova with help of Zamira Xejchieva and Galina Cabdy`rova, and then glossed and analysed by Vlada Baranova.

2.3. Corpus size

The corpus contains 55 texts, 2,076 sentences, and 19,742 tokens. The total duration of the audio recordings is 4 hours and 23 minutes.

2.4. Naming conventions

2.4.1. Folder structure and filenames

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

Folders with text transcripts, organized by genre:

- "conv" (conversations)
- "flk" (folklore texts)
- "nar" (narrative texts)

Each of these genre folders contains one further subfolder per text ("communication"), named identically to the text name. Each text folder contains one or several files with different extensions according to the file type:

- annotated transcript in EXMARaLDA EXB and EXS formats (*.exb, *_s.exs)
- annotated transcript converted into ELAN format (*.eaf)
- annotated transcript converted into ISO/TEI format (*_tei.xml)
- sound file with the recording in WAV format (*.wav) ["standard" package] or MP3 format (*.mp3) ["mp3" package]

Annotated transcripts and original audio files have the file names identical to the text name, except for "_s" and "_tei" suffixes.

Supplementary folders:

- "documentation" (contains the present document)
- "corpus-utilities" (contains annotation panel files that can be opened in EXMARaLDA Partitur Editor):
    - "annotation-panel-inel.xml": annotation values (along with short descriptions) that could be used in tiers SeR, SyF, BOR, CS, IST (in the current version of the corpus these tiers are not annotated).
    - "gloss-panel-kalmyk.xml": annotation values used in the part-of-speech tier (**ps)** and glossing labels for grammatical meanings used in tiers **ge**, **gr**.

Individual files:

- "kalmyk.coma"
- "coma_overview.html" (a browser-readable overview of the main metadata file)

2.4.2. Speaker codes

The speaker codes are derived from the speaker's full names in the order "Family name — First name — Patronymic" in their Latin transliteration. If the name or part of the name is unknown, it is replaced by the letter **N**.

2.4.3. Transliteration of Cyrillic names

Most personal names and placenames in respective metadata fields (except "Region" and "Country") are transliterated from Cyrillic into Latin alphabet following the transliteration standard GOST 7.79 System B (published as GOST 2001)[9]. Exceptions are made if a preferred name spelling exists and for well-known places (e.g. Siberia, Sakhalin).

[9] To transliterate the Cyrillic letter "ц" into Latin, "cz" is recommended in GOST when not before "i", "e", "y", "j"; INEL uses "c" everywhere instead.

Elsewhere, e.g. in text titles, English glosses (**ge** tier) and free translations (**lte** tier), English-style romanization is used.

2.5. Technical formats

2.5.1. Transcripts

The transcripts in the corpus are provided in several formats, all of them in XML. The main working format is EXMARaLDA EXB, while the other formats are derived from EXB to provide search functionalities and alternative ways of access to the data.

EXMARaLDA EXB and EXS

The annotated transcripts are delivered in the formats of the EXMARaLDA software suite. 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".

Please refer to EXMARaLDA documentation for introduction to the use of this software: https://exmaralda.org/en/quickstart-documents/.

Time alignment (synchronization)

The transcripts in the corpus are time-aligned with the available sound recordings. Please be aware that the time alignment is only valid at sentence level (**ts** tier). Technically, time values are also present at word level (**tx** tier), however they should be disregarded as arbitrary. Time values are also technically present in transcripts without any available sound; these are completely arbitrary and should likewise be disregarded.

ELAN EAF

Additionally, the annotated transcripts are converted into ELAN format (".eaf"), which makes the downloaded corpus also browsable and searchable locally using ELAN.

ELAN transcripts differ from the original EXB transcripts in tier structure due to inherent differences between the two data models. In EXB transcripts, the main transcription tier is the tier **tx** (with subdivision into words), and all other tiers are dependent on. In ELAN transcripts, the main transcription tier is the tier **ts** (sentence-level), and all other tiers are dependent on **ts**. Furthermore, annotations on each dependent tier are time-aligned independently of the other tiers, therefore in case of modification of time-alignment and/or merging or splitting annotations the initial alignment between tiers could be broken.

Please be aware that the ELAN versions of the transcripts are provided for compatibility only and are not specifically tested or curated.

ISO/TEI XML

ISO/TEI is an ISO standard (ISO 24624:2016 "Language resource management — Transcription of spoken language"[10]) for representation of spoken data, and at the same time a TEI[11] compliant XML format. It is used, among other, as a source format for the Tsakorpus platform which provides online search over INEL corpora.

2.5.2. Metadata

The corpus metadata are created in the EXMARaLDA Coma (Corpus Manager) and stored in the Coma XML format (file extension ".coma"). One file holds the metadata for the whole corpus.

2.5.3. Media

For texts with audio sources, sound files are provided in Linear PCM WAV format (file extension ".wav"). MP3 versions of all sound files are also provided as a light-weight option (44.1kHz, 192kbps).

2.6. Metadata for the corpus

2.6.1. General corpus metadata

The general metadata about the whole corpus include the corpus name ("INEL Kalmyk Corpus") and some basic metadata fields complying with the standards of DC (Dublin Core) and OLAC (Open Language Archive Community).

[10] https://www.iso.org/standard/37338.html, last access: 07.07.2025
[11] https://tei-c.org/, last access: 07.07.2025

2.6.2. Text (“communication”) metadata

Name: The code which is assigned to the text

Description (Communication):

- **0a Title:** Short title (in English)
- **0b Title (RU):** Short title (in Russian)
- **1 Genre:** Abbreviation of the genre of the text (conv = conversation, flk = folklore, nar = narrative).
- **2a Recorded by**: Person by whom the text was recorded
- **2b Year of recording:** Here the year of recording is given.
- **3 Dialect:** Kalmyk dialect used by the speaker(s) is given here.
- **4 Speakers:** Code(s) of the speaker(s)
- **5 Transcribed by:** Code(s)/Names of the person(s) who did the transcription.
- **6a Translation into Russian:** Code(s)/Names of the person(s) who did the first available translation into Russian
- **6b Translation into English:** All texts were translated with the DeepL software and only partly edited.
- **7 Glossed by:** Code(s)/Names of the person(s) who did the morphological glossing
- **8a-e Annotation SeR / SyF / IST / BOR / CS:** These tiers were not annotated in the current version of the corpus.

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

- **1 Country:** All the texts were recorded in Russia.
- **2 Region:** All the texts were recorded in the Republic of Kalmykia
- **3a Settlement:** The place of the recording
- **3b Settlement (RU):** The place of the recording in Russian
- **3c Settlement (LatLng):** Geographic coordinates (latitude, longitude) of the settlement

Languages:

- **Language code:** The ISO-code of the language of the text. It is always “xal” – Kalmyk.

Setting:

- **Has audio:** Marked “yes” if a sound recording is available, otherwise marked “no”.

Recording: If an audio file is available, it is linked to the text description.

Transcriptions: The basic transcription (.exb) and the segmented transcription (_s.exs) are linked here to the text description; the latter is needed for searching the corpus.

Attached file(s): All other relevant files are linked to the text description here if available.

2.6.3. Speaker metadata

Metadata of speaker(s) include primarily biographical information of the speaker. Name fields exist both in Russian (RU) and English. The following fields are defined:

Sigle: Speaker code as defined in 2.4.2

Pseudo: Name shown in Coma’s main view (using family name, first name and patronymic)

Sex: male or female

Description:

- **1a-b Family name (EN, RU)**
- **2a-b Given name (EN, RU)**
- **3a-b Patronymic (EN, RU)**
- **4a-b Maiden name (EN, RU):** Maiden name is given here.

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

- **1a-b Place of birth (EN, RU)**
- **1c Place of birth (LatLng)**: Geographic coordinates (latitude, longitude) of place of birth
- **2 Region**
- **3 Country**
- **4 Year of birth**

- **5a-b Domicile (EN, RU):** The current (i.e. at the time of the recording) place of residence of the speaker if known
- **5c Domicile (LatLng):** Geographic coordinates (latitude, longitude) of the current residence

Occupation: Here information – if available – is given on the speaker's occupation/profession.

- **1a-b Occupation (EN, RU)**

Ethnicity:

- **Ethnicity:** always "Kalmyk"

L1 (Language): Here information about the first language of the speaker is given.

- **Language code:** "xal" – Kalmyk or "rus" – Russian.
- **First language:** The name of the language

3. Transcription and annotation

The general principles of transcription, annotation and translation in many respects go back to the Nganasan Spoken Language Corpus, documented in the respective user guidelines (Wagner-Nagy et al. 2018). They largely follow those of the INEL project as described in Arkhipov (2020).

3.1. Tier layout

Every annotation tier has a distinct label shown in EXB files.

Table 1. Overview of annotation tiers used in EXMARaLDA-transcripts[12]

Tier label	Tier name	Description	Unit	Optionality
ref	Reference	Text ID + sentence number Text ID + speaker code + sentence number (for texts with multiple speakers)	sentence	obligatory
ts	Text (sentence)	Main transcription	sentence	obligatory
tx	Text (word)	Main transcription segmented by word	word	obligatory
mb	Morpheme breaks	Morpheme breakdown of words	morph	obligatory
mp	Morphemes (lexical)	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
fr	Free translation (Russian)	Free translation (Russian)	sentence	obligatory
lte	Automatic translation (English)	Automatic translation into English produced with the DeepL software[13] from **fr**-tier	sentence	obligatory
nt	Notes	Notes (in English or in Russian)	sentence	optional
SeR	Semantic Role	Semantic (thematic) roles of NPs	word	absent
SyF	Syntactic function	Syntactic functions of predicates and arguments, as well as for subordinate clauses	word	absent
BOR	Borrowing	Borrowings (source language and borrowing type)	word	absent

[12] There are also empty tiers **fe** and **fg** in the transcripts.

[13] https://www.deepl.com

Tier label	Tier name	Description	Unit	Optionality
CS	Code-switching	Code-switching and calques (source language and type)	group of words	absent
IST	Information status	Information status for major NPs (given/new/accessible)	word	absent
fe	Free translation (English)	Free translation (English)	sentence	absent
fg	Free translation (German)	Free translation (German)	sentence	absent

The tiers **SeR**, **SyF**, **BOR**, **CS, IST, fe** and **fg** are not annotated in the current version of the corpus.

Figure 1 gives an example of how a sentence looks like in the corpus (empty tiers are omitted):

Figure 1. *A sample transcript fragment showing the tier layout*

ref	XZD_2008_Fool_flk.002				
ts	Kezänä bääǯə emgən övgən bääǯə.				
tx	Kezänä	bääǯə	emgən	övgən	bääǯə.
mb	kezänä	bää-ǯ	emgə-n	övgə-n	bää-ǯ
mp	kezänä	bää-ǯə	emgə-n	övgə-n	bää-ǯə
ge	formerly	be-CVB.IPFV	old.woman-EXT	old.man-EXT	be-CVB.IPFV
gr	раньше	быть-CVB.IPFV	старуха-EXT	старик-EXT	быть-CVB.IPFV
mc	adv	v-v:(conv)	n-n:(case)	n-n:(case)	v-v:(conv)
ps	adv	v	n	n	v
fr	Жили-были старик и старуха.				
fe	Once upon a time there was an old man and an old woman.				

3.2. Annotation tiers

3.2.1. Reference (ref)

The reference tier (**ref**) for each sentence contains the text name and the number of the sentence, separated by a full stop. The sentences are numbered throughout the entire text. The sentence numbers are zero-padded up to 3 digits (see Figure 1). This part of the **ref** tier should be used for citation of a specific sentence coming from the corpus. In texts recorded from multiple speakers, the speaker code is additionally provided between the text code and the sentence number, separated by dots. The numbering is consecutive within each speaker separately, starting from 001.

3.2.2. Morpheme breaks (mb)

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

3.2.3. Morphemes (morphophonological) (mp)

The underlying morphemes tier (mp) shows the lexical representation of the morphs, both stems and affixes, which appear in the mb tier. It follows the lexicon of *SIL Fieldworks Language Explorer* (FLEx),[14] where the texts were glossed.

3.2.4. Gloss (ge, gr)

The gloss tiers (ge, gr) contain the English and Russian glossing of the morphemes in mb and mp. Stems receive their respective lexical glosses in the two languages, while affixes are glossed identically in capital Latin letters and mostly

[14] https://software.sil.org/fieldworks/, last access: 07.07.2025.

according to the Leipzig Glossing Rules.15 If the grammatical meaning is absent in the Leipzig Glossing Rules, the English abbreviation of the most commonly used term, accepted in INEL and corresponding to the Mongolian tradition, was used. If a morpheme contains two or more semantic components, then they are separated by a dot. For the list of abbreviations used see Appendix A1.

3.2.5. Morphological category (mc)

The **mc** tier indicates the morphological category of both lexical stems (i.e. the part of speech) and affixes (i.e. the inflectional category or the derivational process). Table 2 and Table 3 show the tags used for lexical stems, inflectional categories and other affixes. Tags for inflectional categories are marked as *x:(cat)*, where *x* is the corresponding lexical stem tag, *cat* is a tag for the category filling an optional slot (e.g. *n:(case)* – case of nominals). Derivational processes are marked as *x>y*, *x* and *y* being the tags for lexical stems (e.g. *n>v* for verbalizers deriving verbs from nominals).

Table 2. Tags for lexical stems

Tag	**Description**
adj	adjective
adp	adposition/postposition
adv	adverb
conn	connective
exist	existentional copula
ideo	ideophone
interj	interjection
n	noun
num	numeral
pron	pronoun
ptcl	particle
pers	personal pronoun
ptcp	participle (used only in derivational processes)
v	verb

Table 3. Tags for inflectional categories

Tag	**Comment**
Inflection of nominals	
n:(case)	case suffix on nouns (also on adjectives, participles and pronouns)
n:(num)	number suffix on nouns (also on adjectives and pronouns)
n:(poss)	possessive suffix on nouns (also on adjectives, numerals, participles and pronouns)
n:(poss.p.num)	reflexive suffix on nouns (also on adjectives, participles and pronouns)
Inflection of verbs	
v:(arg)	changing the argument structure of verbs (causative, passive, reciprocal suffix, etc.)
v:(asp)	aspect
v:(conv)	converb suffix on verbs
v:(inter)	interrogative clitic
v:(mood)	mood suffix (imperative, jussive, permissive, dubitative and evidentiality)
v:(neg)	negation suffix/clitic
v:(pn)	person-number suffix on verbs

[15] https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf, last access: 07.07.2025.

Tag	Comment
v:(pn.poss)	reflexive suffix on verbs with combination of person-number
v:(poss)	possessive suffix on participles
v:(tense)	tense suffix on verbs
v:(tense.pn)	cumulative tense and person-number marker
Inflection of participles	
ptcp:(NEG)	negation suffix attached to participles

3.2.6. Part of speech (ps)

The part of speech tier (**ps**) contains information about the grammatical category of each word form. Hence, the outcome of derivational processes is marked here. The tags used here are slightly different from those used in the morphological category tier **mc** (see Table 2). Personal pronouns (**mc**: *pers*) are tagged as pronouns (*pron*) in the **ps** tier. Participles (**mc**: *ptcp*), as well as converbs, are tagged as verbs (*v*).

3.2.7. Free translation (fr)

The free translation ties **fr** gives free translation of the utterance in question into Russian.

3.2.8. Automatic English translation (lte)

The tier **lte** contains the automatic translation into English produced with the DeepL software (partly with editing).

3.2.9. Notes (nt)

The Notes tier (**nt**) contains notes related to the sentence.

3.3. Searching the corpus

3.3.1. Search with EXMARaLDA EXAKT

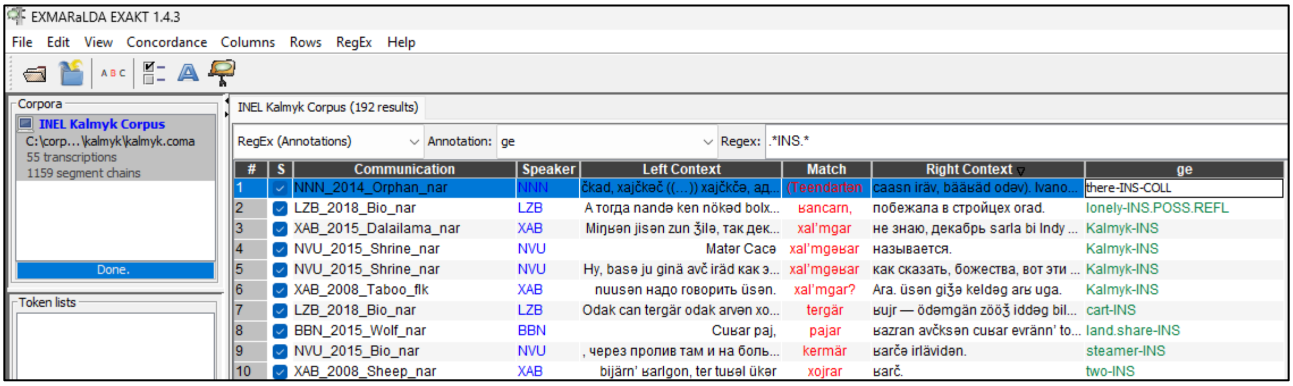
The EXMARaLDA software suite includes EXAKT, an analysis and concordance tool.

In order to perform a search on the downloaded corpus files locally, the main metadata file (**kalmyk.coma**) should be opened with “File > Open Corpus” command. (Creating a word list is optional.)

One of the tiers should be selected in the main concordance window: either one of the annotation tiers (recommended; use “RegEx (Annotations)”; select any of tiers except **tx** under “Annotation”) or the transcription tier (**tx**; use “RegEx (Transcription)”).

- A search expression (interpreted as a regular expression[16]) should be specified in the **Regex** field. The matching results will be displayed in a column corresponding to the selected tier, e.g. “ge”. Please refer to section 3.2 and Appendix A1 for annotations used in the corpus.
- Note that only the part matching the search expression will be displayed in the column. E.g. when searching for an instrumental case marker with “INS” in tier **ge**, only “INS” will be shown in the “ge” column. In order to have the complete word gloss displayed in the “ge” column, enter “.*INS.*” as search expression.

Figure 2. *EXAKT search window*



INEL Kalmyk Corpus
C:\corp…\kalmyk\kalmyk.coma
55 transcriptions
1159 segment chains

INEL Kalmyk Corpus (192 results)

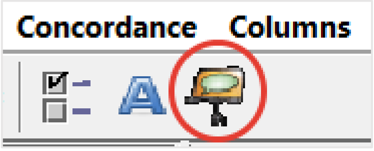
RegEx (Annotations) Annotation: ge Regex: .*INS.*

#	S	Communication	Speaker	Left Context	Match	Right Context	ge
1	☑	NNN_2014_Orphan_nar	NNN	čkad, xajčkəč ((…)) xajčkčə, ад...	(Teendartən	caasn iräv, bääʁäd odəv). Ivano...	there-INS-COLL
2	☑	LZB_2018_Bio_nar	LZB	А тогда nandə ken nökəd bolx...	ʁancarn,	побежала в стройцех orad.	lonely-INS.POSS.REFL
3	☑	XAB_2015_Dalailama_nar	XAB	Miŋʁən jisən zun ǯilə, так дек...	xal’mgar	не знаю, декабрь sarla bi Indy ...	Kalmyk-INS
4	☑	NVU_2015_Shrine_nar	NVU	Matər Cacə	xal’mgəʁar	называется.	Kalmyk-INS
5	☑	NVU_2015_Shrine_nar	NVU	Ну, basə ju ginä avč iräd как э...	xal’mgəʁar	как сказать, божества, вот эти ...	Kalmyk-INS
6	☑	XAB_2008_Taboo_flk	XAB	nuusən надо говорить üsən.	xal’mgar?	Ага. üsən giǯə keldəg arʁ uga.	Kalmyk-INS
7	☑	LZB_2018_Bio_nar	LZB	Odak can tergär odak arvən xo...	tergär	ʁujr — ödəmgän zööǯ iddəg bil...	cart-INS
8	☑	BBN_2015_Wolf_nar	BBN	Cuʁar paj,	pajar	ʁazran avčksən cuʁar evränn’ to...	land.share-INS
9	☑	NVU_2015_Bio_nar	NVU	, через пролив там и на боль...	kermär	ʁarčə irlävidən.	steamer-INS
10	☑	XAB_2008_Sheep_nar	XAB	bijärn’ ʁarlgon, ter tuʁəl ükər	xojrar	ʁarč.	two-INS

[16] https://www.regular-expressions.info/, last accessed 28.04.2025.

- The “Match” column represents the content of the **tx** tier (word or sentence) corresponding to the annotation found in the specified tier. Double-click the entry in the “Match” column to display a portion of the entire transcript containing the example found (all tiers) in the lower part of the screen. After that, a click on the “Open Partitur” button will open the entire transcript in EXMARaLDA Partitur Editor.

Figure 3. *EXAKT: “Open Partitur” button*



Please refer to EXMARaLDA manuals[17] for further details on using EXAKT and Partitur Editor.

3.3.2. Online search in Tsakorpus

Online search in the corpus is provided via Tsakorpus, an open-source search platform for linguistic corpora. The current version of the corpus can be accessed at https://inel.corpora.uni-hamburg.de/KalmykCorpus/search. The interface of online search is available in English and in Russian.

Tsakorpus offers the following possibilities:

- Search in multiple annotation tiers
- Search for substring, simple patterns (using *) or regular expressions
- Multi-word search (with or without distance restrictions)
- Negative queries (sentences which do NOT have a word with specified parameters)
- Search for sentences, words (wordforms), lemmas
- Search in a subcorpus
- Exporting search results as CSV/XLSX

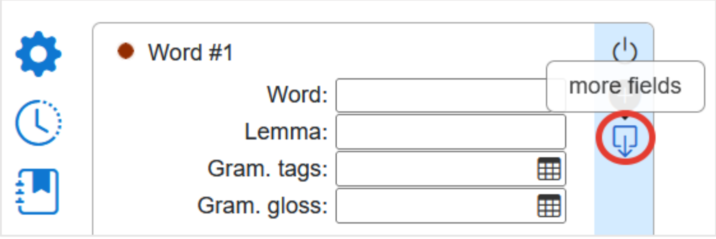
To run a search in the main transcription tier (**tx**) or in the word- and morph-level annotation tiers, “Language/tier” field should be set to “Kalmyk” and the search expression(s) entered in one or more corresponding fields.

Table 4. Tsakorpus search fields and EXMARaLDA tiers: main transcription and word-/morph-level annotation

Tsakorpus search field	**Corresponding tier in EXMARaLDA**
Word	**tx**
Lemma	**mp** (stem)
Gram. tags	**ps**; grammar tags generated from grammatical glosses (**ge, gr**)
Gram. gloss	grammatical (i.e. affix) glosses (**ge, gr**)
Lex. gloss (en)*	lexical (i.e. stem) glosses (**ge**)
Lex. gloss (ru)*	lexical (i.e. stem) glosses (**gr**)
Morph. slot*	**mc**
Part of speech (syntax)*	**ps**

*To display search fields marked with *, click on “More fields” button next to “Word” and “Lemma” fields.

Figure 4. *Tsakorpus interface: Show more fields*



Please refer to section 3.2 and Appendix A1 for annotations used in the corpus.

[17] https://exmaralda.org/en/quickstart-documents/, last accessed: 28.04.2025.

Lexical and grammatical glosses in Tsakorpus

Each word in Tsakorpus is internally split into stems (lexical items) and affixes (grammatical morphs).

The stem can be found by searching for its underlying (**mp**) form (e.g. “kür”) in the **Lemma** field, or by searching for its lexical gloss (e.g. “reach” / “достигать”) in **Lex. gloss (en)** or **Lex. gloss (ru)** fields.

The affixes can be found by searching for the complete gloss (e.g. “PST.1PL”) in the **Gram. gloss** field, or with corresponding grammar tags (e.g. “pst,pn1,pnpl”) in the **Gram. tags** field (see next section for details on grammar tags).

To find only a particular allomorph, its form can be specified in curly braces following the gloss in the **Gram. gloss** field: “PST.1PL{üdn}”.

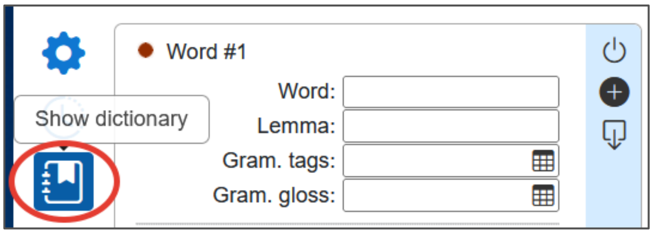
In case there exist more than one underlying form of an affix in **mp** tier, a particular underlying form can be specified in curly braces with underscore in the **Gram. gloss** field: “NEG{_go}”.

Table 5. Stems and affixes in Tsakorpus

EXMARaLDA tier	**Word**	**Stem**	**Search field**	**Affix**	**Search field**
tx	kürüdn				
mb	kür-üdn	kür		üdn	Gram. gloss: PST.1PL{ üdn }
mp	kür-udən	kür	Lemma: kür	udən	Gram. gloss: PST.1PL{_ udən}
ge	reach-PST.1PL	reach	Lex. gloss (en): reach	PST.1PL	Gram. gloss: PST.1PL Gram. tags: pst,pn1,pnpl
gr	достигать-PST.1PL	достигать	Lex. gloss (ru): достигать	PST.1PL	Gram. gloss: PST.1PL Gram. tags: pst,pn1,pnpl
ps	v		Part of speech (syntax): v		

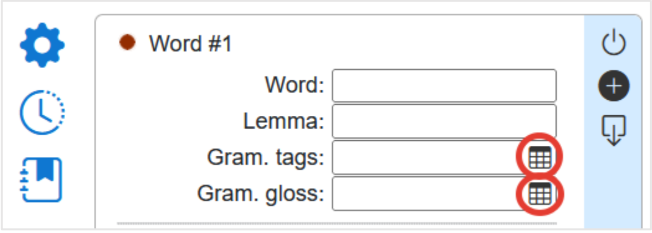
A list of lemmas (i.e. underlying forms of stems as represented in **mp** tier) along with their translations (lexical glosses) can be displayed with “Show dictionary” button.

Figure 5. *Tsakorpus interface: Show dictionary*



For most word- and morph-level annotation tiers, such as grammar tags, grammatical glosses, borrowings, one can either type in the search expression directly or choose from the list of available values. To open the list of values, click on the icon in the search field.

Figure 6. *Tsakorpus interface: Show list of values*



Grammatical glosses and grammar tags in Tsakorpus

In addition to grammatical glosses as present in tiers **ge**, **gr**, Tsakorpus provides another search possibility called “grammar tags.” Grammar tags are generated by rules based on part of speech and glosses. For a complete list of glosses and grammar tags please refer to Table 8 (Appendix A1).

- Tags are assigned to an entire word and not to a particular morpheme in a word.
- By default, grammar tags are identical to a lower-case version of the corresponding gloss or part of speech label, e.g. (part of speech) “v” => “v”, (gloss) “PL” => “pl”. Exceptions are mostly due to avoiding overlapping.

- Stems with glossing labels similar to a grammatical gloss, e.g. “NEG.COP” for “existential negation”, will also be assigned grammar tags. Such glosses are marked as “lexical” in Comments columns in **Table 8** (Appendix A1). They can be found with either **Gram. tags** or **Lex. gloss (en)** / **Lex. gloss (ru)** fields, but not with **Gram. gloss** field.
- When a gloss consists of multiple components, such as “PST.1PL”, each of them is usually translated into a tag, e.g. “PST.1PL” => “pst” (past tense), “pn1” (1 person), “pnpl” (plural number). A search for tag “pn1” will return all words with any of glosses “1SG”, “1PL”, “PRO1SG”, “PST.1PL” etc.
- When searching with glosses, the entire gloss should be entered as a search expression. E.g. a search for “PST” will not find “PST.1PL”. Use grammar tags if you need to search for a component of a complex gloss.
- When specifying more than one tag in a search expression, they can be combined with logical operators: AND (“,”), OR (“|”) and NOT (“~”), e.g. “pst,pn1,pnpl”. When selecting tags from the list of values, multiple tags which are listed as belonging to the same Tsakorpus category (see **Table 8** in Appendix A1) will be by default joined by OR (“|”), e.g. “(acc|dat)”. Multiple tags which are listed as belonging to different Tsakorpus categories will be by default joined by AND (“,”), e.g. “acc,pl”.

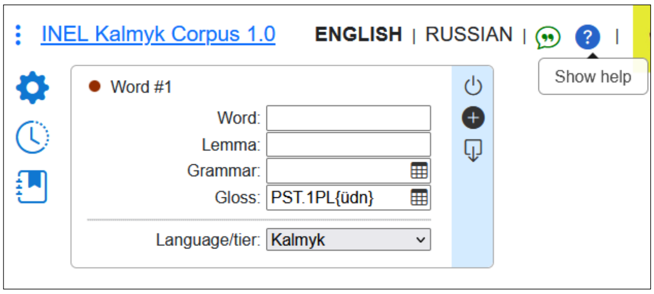
To search in one of the sentence-level annotation tiers, the search expression should be entered into “Word” field, and “Language/tier” field set to one of the following:

Table 6. Tsakorpus search fields and EXMARaLDA tiers: sentence-level annotation

Language/tier label	**Corresponding tier in EXMARaLDA**
Russian	**fr**
English	**lte**
Note	**nt**
ID	**ref**

For further details please refer to Tsakorpus online help.

Figure 7. *Tsakorpus interface: Show help*



Word:

Lemma:

Grammar:

Gloss: PST.1PL{üdn}

Language/tier: Kalmyk

References

Arkhipov, Alexandre. 2020. *INEL corpora general transcription and annotation principles*. Szeged – Hamburg: University of Szeged & University of Hamburg. (Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology 5.) https://doi.org/10.14232/wpcl.2020.5

Baranova, Vlada & Say, Sergey. 2009. Kommentarii k korpusu tekstov [Note on texts]. *Issledovanija po grammatike kalmyckogo jazyka. Acta Linguistica Petropolitana* 5(2): 710-729.

GOST 2001 — GOST 7.79–2000. *Sistema standartov po informacii, bibliotečnomu i izdatel’skomu delu. Pravila transliteracii kirillovskogo pis’ma latinskim alfavitom* [GOST 7.79–2000.The system of standards in information, librarianship and publishing. Rules of transliteration of Cyrillic letters into the Latin alphabet]. Minsk: Mežgosudarstvennyj sovet po standartizacii, metrologii i sertifikacii. https://protect.gost.ru/document.aspx?control=7&id=130715, https://ifap.ru/library/gost/7792000.pdf

Janhunen, Juha. 2006. Mongolic languages. In: Brown, K. (ed.) *The encyclopedia of language & linguistics*. (231–234). Amsterdam: Elsevier.

Rákos, Attila. 2015. *Synchronic and Diachronic Comparative Analysis of the Oirad Dialects*. Doktori Disszertacio. Budapest: Eotvos Lorand Tudomanyegyetem, Bolcseszettudomanyi Kar.

Texts 2009. Texts. *Issledovanija po grammatike kalmyckogo jazyka. Acta Linguistica Petropolitana* 5(2): 730-856.

Wagner-Nagy, Beáta & Szeverényi, Sándor & Gusev, Valentin. 2018. *User’s guide to Nganasan Spoken Language Corpus*. Szeged – Hamburg: University of Szeged & University of Hamburg. (Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology 1.) https://doi.org/10.14232/wpcl.2018.1.

Appendix A1. Morpheme glossing labels (tiers ge, gr) and Tsakorpus grammar tags

Table 7 presents a list of grammatical glosses used in tiers **ge**, **gr**, sorted alphabetically. Table 8 lists the glossing labels (for affixes and lexemes) and corresponding grammar tags for use in Tsakorpus online search. It is sorted by category.

Table 7. Morpheme glossing labels

Label	**Description**
1	first person
2	second person
3	third person
ABL	ablative case
ACC	accusative case
ACT	active participle
AFF	affirmative particle
ANT	anterior converb
CAUS	causative
COLL	collective
COM	comitative case
COMPL	completive
CONC	concessive converb
COND	conditional converb
CONT	continuative participle
COP	copula
CVB	converb
DAT	dative case
DIR	directive case
DIST	distributive numeral
DUB	dubitative
EMPH	emphatic
EVD	evidentiality
EXT	extension (the unstable nasal *-n* in some nouns)
FUT	future
GEN	genitive case
HAB	habitual
IMP	imperative mood
%IMP	another form of imperative
INS	instrumental case
INTR	intransitive verb
IPFV	imperfective aspect
JUSS	jussive
MIR	mirative
MOD	modal converb
NEG	negation
NMLZ	nominalization
ORD	ordinal numerals
PASS	passive
PERM	permissive
PL	plural
PLR	verbal plurality, pluractionality
POSS	possessive
PROG	progressive
PROH	prohibitive
PROP	proprietive case
PRS	present tense
PST	past tense
PTCL	particle
PTCP	participle

Label	Description
PURP	purposive converb
Q	question particle
Q.NP	question to noun
RDPL	reduplication
RECP	reciprocal
REFL	reflexive
REM	remote past tense (“past temporal frame”)
SIM	similative
SG	singular
SOC	sociative causative
TERM	terminative converb
VBLZ	verbalizer

Table 8. List of morpheme glossing labels by category

Tag	Description	Tsakorpus grammar tags	Tsakorpus category	Comment
Person and number				
1SG	1 person, singular	pn1,pnsg	pers,pnum	
1PL	1 person, plural	pn1,pnpl	pers,pnum	
2SG	2 person, singular	pn2,pnsg	pers,pnum	
2PL	2 person, plural	pn2,pnpl	pers,pnum	
3SG	3 person, singular	pn3,pnsg	pers,pnum	
3PL	3 person, plural	pn3,pnpl	pers,pnum	
3	3 person	pn3	pers	only as a tag
Nominal categories				
Case				
EXT	extension	ext	n-case	
DAT	dative case	dat	n-case	
ACC	accusative case	acc	n-case	
PROP	proprietive case	prop	n-case	
GEN	genitive case	gen	n-case	
ABL	ablative case	abl	n-case	
INS	instrumental case	ins	n-case	
COM	comitative case	com	n-case	
DIR	directive case	dir	n-case	
NOM	nominative case	nom	n-case	only with pronouns
Number				
PL	plural	pl	num	
COLL	collective	coll	num	
Possessives				
POSS	possessive	poss	poss	only in combinations
POSS.REFL	possessive, reflexive	poss,refl	poss,misc	
POSS.1SG	possessive, 1 person, singular	poss,pn1,pnsg	poss,pers,pnum	
POSS.1PL	possessive, 1 person, plural	poss,pn1,pnpl	poss,pers,pnum	
POSS.2SG	possessive, 2 person, singular	poss,pn2,pnsg	poss,pers,pnum	
POSS.2PL	possessive, 2 person, plural	poss,pn2,pnpl	poss,pers,pnum	
POSS.3	possessive, 3 person	poss,pn3	poss,pers	

Tag	Description	Tsakorpus grammar tags	Tsakorpus category	Comment
Verbal categories				
Argumental structure				
CAUS	causative	caus	v-arg	
RECP	reciprocal	recp	v-arg	
PASS	passive	pass	v-arg	
PLR	verbal plurality, pluractionality	plr	v-arg	
SOC	sociative causative	soc	v-arg	
Tense				
PRS	present tense	prs	v-tense	
PST	past tense	pst	v-tense	
PST.1SG	past tense, 1 person, singular	pst,pn1,pnsg	v-tense,pers,pnum	
PST.1PL	past tense, 1 person, plural	pst,pn1,pnpl	v-tense,pers,pnum	
REM	remote past tense	rem	v-tense	
Aspect				
PROG	progressive	prog	v-asp	
COMPL	completive	compl	v-asp	
CONT	continuative participle	cont	v-asp	
Mood				
COP.MIR	copula, mirative	cop.mir	v-mood	
DUB	dubitative	dub	v-mood	
EVD	evidentiality	evd	v-mood	
JUSS	jussive	juss	v-mood	
PERM	permissive	perm	v-mood	
Imperative				
IMP	imperative	imp	v-imp	
IMP.PL	imperative, plural	imp,pl	v-imp,num	
IMP.1PL	imperative, 1 person, plural	imp,pn1,pnpl	v-imp,pers,pnum	
%IMP	possible imperative	%imp	v-imp	
PROH	prohibitive	proh	v-imp	lexical
Converbs				
CVB.ANT	anterior converb	cvb.ant	cvb	
CVB.IPFV	imperfective converb	cvb.ipfv	cvb	
CVB.COND	conditional converb	cvb.cond	cvb	
CVB.MOD	modal converb	cvb.mod	cvb	
CVB.PURP	purpose converb	cvb.purp	cvb	
CVB.TERM	terminative converb	cvb.term	cvb	
CVB.CONC	concessive converb	cvb.conc	cvb	
Participles				
PTCL.CONC	concessive participle	ptcl.conc	ptcp	
PTCP.HAB	habitual participle	ptcp.hab	ptcp	
PTCP.FUT	future participle	ptcp.fut	ptcp	
PTCP.PST	past participle	ptcp.pst	ptcp	
PTCP.FUT.COP	future participle, copula	ptcp.fut.cop	ptcp	

Tag	Description	Tsakorpus grammar tags	Tsakorpus category	Comment
PTCP.ACT	active participle	ptcp.act	ptcp	
PTCP.PASS	passive participle	ptcp.pass	ptcp	
Derivations				
VBLZ	verbalizer	vblz	drv	
ORD	ordinal numbers	ord	drv	
DIST	distributive numeral	dist	drv	
NMLZ	nominalization	nmlz	drv	
Clitics				
COP.AFF	copula, affirmative particle	cop.aff	clt	
PTCL.AFF	affirmative particle	ptcl.aff	clt	
Q	question particle	q	clt	
Q.NP	question to noun	q.np	clt	
Miscellaneous				
EMPH	emphatic	emph	misc	
REFL	reflexive	refl	misc	
RDPL	reduplication	rdpl	misc	
NEG	negation	neg	misc	
NEG.COP	existential negation	neg.cop	misc	lexical
Personal pronouns				
PRO1SG	personal pronoun, 1 person singular	pn1,pnsg	pers,pnum	lexical
PRO1PL	personal pronoun, 1 person plural	pn1,pnpl	pers,pnum	lexical
PRO2SG	personal pronoun, 2 person singular	pn2,pnsg	pers,pnum	lexical
PRO2SG.RESP	personal pronoun, 2 person singular (respectful)	pn2,pnsg,resp	pers,pnum,resp	lexical
PRO2PL	personal pronoun, 2 person plural	pn2,pnpl	pers,pnum	lexical
PRO3SG	personal pronoun, 3 person singular	pn3,pnsg	pers,pnum	lexical
PRO3PL	personal pronoun, 3 person plural	pn3,pnpl	pers,pnum	lexical