Alexandre Arkhipov

INEL Corpora General Transcription and Annotation Principles

Working Papers in Corpus Linguistics
and Digital Technologies:
Analyses and Methodology
Vol. 5.

Alexandre Arkhipov

INEL Corpora General Transcription and Annotation Principles

Working Papers in Corpus Linguistics and Digital Technologies:

Analyses and Methodology

Vol. 5.

Szeged – Hamburg

2020

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

Vol. 5

WPCL issues do not appear according to strict schedule.

Vol. 5 (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 978-963-306-778-9 (pdf)
DOI: https://doi.org/10.14232/wpcl.2020.5

Contents

1. Introduction 3
   1.1. The INEL corpora: objectives and data 3
   1.2. The INEL corpora: data formats and workflows 4
2. Segmentation and transcription: general principles 4
   2.1. Main transcription tiers in the INEL corpora layout 4
   2.2. Segmentation levels: sentences, words and morphemes 5
      2.2.1. Major units: sentences 5
      2.2.2. Punctuation: orthography-based 7
      2.2.3. Words and morphemes 7
   2.3. General transcription principles 8
      2.3.1. Elementary units: Phonological/broad phonetic 8
      2.3.2. Characters: Non-IPA 9
      2.3.3. Alternative transcriptions 9
   2.4. Speakers, turn-taking and direct speech 10
3. Translations and comments: general principles 11
   3.1. Main translation and comments tiers 11
   3.2. Alternative translations and comments 12
4. Transcription and translations: special cases 12
   4.1. Uncertainty in form and meaning 12
      4.1.1. Uncertainty 12
      4.1.2. Unintelligible fragments 14
      4.1.3. Unclear sentences 14
   4.2. Disfluencies and non-speech events 14
      4.2.1. Non-speech sounds and events 14
      4.2.2. False starts/self-repairs 15
      4.2.3. Annotator corrections 15
   4.3. Mismatches between original text and translation 16
      4.3.1. Reconstructed referents 16
      4.3.2. Material added or omitted for grammaticality 16
      4.3.3. Significant material added for clarification 16
      4.3.4. Literal vs. idiomatic translation 17
5. Annotation of borrowings, calques and code-switching 17
   5.1. Borrowings 18
      5.1.1. Source language 18
      5.1.2. Semantic/lexical class of borrowing 18
      5.1.3. Phonological adaptation 20

5.1.4. Morphological adaptation 20
5.2. Code-switching and calques 21
5.2.1. Code-switching 21
5.2.2. Calques 21
References 23
Appendix A. Complete tier layout examples 25
Appendix B. Summary of notation conventions 28

1. Introduction[1]

1.1. The INEL corpora: objectives and data

INEL ("Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages") is a long-term research project (2016–2033), whose primary goal is to create digital annotated corpora of several languages of Northern Eurasia, making possible typologically aware corpus-based grammatical research. For an overview of the project, see [Arkhipov & Däbritz 2018].

As of December 2020, full versions have been released for two corpora, Kamas and Dolgan [Gusev et al. 2019; Däbritz et al. 2019]; two intermediate versions have also been published for Selkup [Brykina et al. 2020], while the complete Selkup corpus is scheduled to appear by the end of 2021. Evenki is another currently running subproject, for which the corpus is also to appear at the same time.

In this paper, we outline the basic principles of transcription and some aspects of other annotations (such as translations and annotations of code switching), common for all the INEL corpora. However, each INEL corpus will have its own specifics which are not covered here (see [Arkhipov et al. 2020] for Kamas and [Däbritz 2020] for Dolgan).

The common system presented here should not be considered a pre-defined departure point, but rather a convergence point for the individual corpora. The varied data sources in which the INEL corpora take their origin, and the differing workflows associated with each of them, implied a great deal of variation in many aspects. As desirable as it might seem to have maximal unification across the corpora, it would sometimes have caused too much delay to elaborate a common decision suitable to all use cases prior to producing annotated data. Aspects reflected in the present document were either applied uniformly since the very beginning, or got unified and standardized over time, and some decisions were not carried out retrospectively to the previously annotated material. One should not forget either that annotation is only a means for asking questions, and changes to the annotation schemes and principles might become desirable as new data and/or new questions come into the light. So ideally annotating a corpus should be an iterative process (cf. [Dickinson & Tufiş 2017] for different understandings of iterative enhancement), although in practice only some smaller fragments of annotation can possibly be adjusted or enhanced corpus-wide after the main body of the work is done.

Many aspects of transcription and annotation in INEL corpora originate from the Nganasan Spoken Language Corpus (see [Wagner-Nagy et al. 2018]), but there have been also a number of changes or additions. The ensemble of the conventions is influenced by three major groups of considerations, often contradictory:

i. linguistic relevance, including a trade-off between packaging more information into a corpus and lowering the complexity for the end user / for the corpus developers;
ii. nature of the primary data and the degree of certainty of the analyses, partly depending on the availability of language consultants who could reliably interpret the data;
iii. technical constraints imposed by the software and the workflows.

Concerning (ii), all the corpora so far are built on heterogeneous data sets. Part of the sources come in written form (either as manuscript fieldnotes, or as published collections of folklore), the other part coming from more or less recent sound recordings, usually previously untranscribed. The combination of these two types of sources is relevant for segmentation and transcription issues discussed in section 2. The next subsection will briefly present the technical considerations mentioned in (iii).

[1] This paper 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.

1.2. The INEL corpora: data formats and workflows

The working formats of the corpora are those of the EXMARaLDA software suite[2] [Schmidt & Wörner 2014], all of them in XML. The main transcript files that can be used for browsing the transcripts with the EXMARaLDA Partitur Editor have 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.

The transcripts are however not originally created in Partitur Editor. The main linguistic analysis, i.e. interlinear glossing and editing, is performed in SIL FLEx[3]. For data from sound recordings, the transcription is first created in ELAN[4] [Sloetjes & Wittenburg 2008] and then imported into FLEx, preserving time alignment at sentence level until the export into EXMARaLDA format. Notably, FLEx imposes its limitations on segmentation, forcing sentences to split at hard-coded set of punctuation characters (see 2.1).[5]

When finalized, the INEL corpora are also exported into ISO/TEI Standard Transcription of spoken language,[6] a tool-independent standard format for spoken data. At present, it serves as input for the online search engine on Tsakorpus platform[7] (see more about application of ISO/TEI to INEL data in [Arkhangelskiy et al. 2019; Ferger & Jettka 2020]). Importantly, the Tsakorpus search operates within ISO/TEI sentences (utterances) and cannot run complex queries across sentence boundaries.

Fig. 1 summarizes the basic data transformation flow:

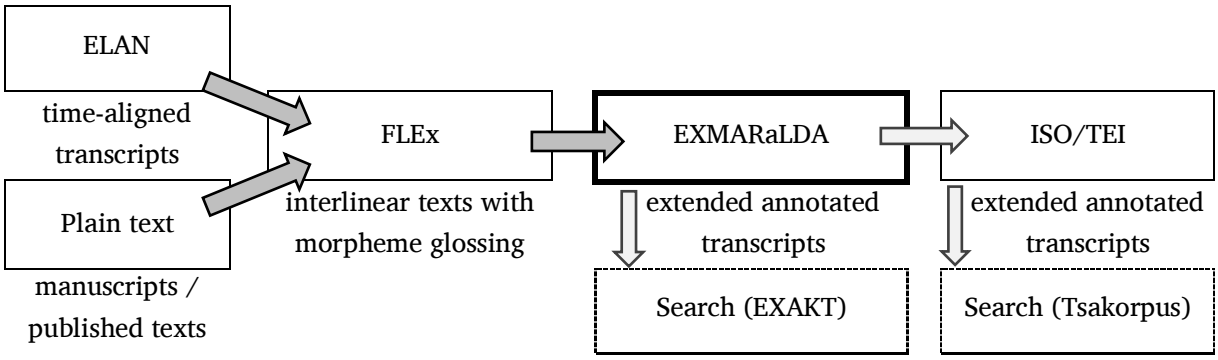

Figure 1. *General data transformation flow in INEL corpora*

2. Segmentation and transcription: general principles

2.1. Main transcription tiers in the INEL corpora layout

The INEL corpora are developed in a multi-tier layout with as many as up to 25 tiers per speaker, some of which are optional. A sample layout is presented in Appendix A below; for a detailed account of the tier system please refer to the documentation of individual corpora, e.g. [Arkhipov et al. 2020: §2.10].

Let us first mention the main transcription tiers, **tx** and **ts** (see (1) below). The difference between them is that **ts** presents transcriptions of entire sentences, while **tx** has the same content divided into

[2] http://exmaralda.org/en/, last access: 25.11.2020.
[3] SIL Fieldworks Language Explorer. https://software.sil.org/fieldworks/, last access: 25.11.2020.
[4] https://archive.mpi.nl/tla/elan, last access: 25.11.2020. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands.
[5] This is especially problematic for imported audio transcripts, since time alignment is lost if a single time-aligned ELAN annotation is split by FLEx based on internal punctuation. Therefore one needs to avoid using annotation-internal punctuation during transcribing or split such annotations manually before export from ELAN into FLEx.
[6] http://www.iso.org/iso/catalogue_detail.htm?csnumber=37338, last access: 10.12.2020.
[7] https://bitbucket.org/tsakorpus/, last access: 25.11.2020.

words. The latter is the basis for the morpheme breakdown 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**. The **tx** tier is thus crucial for import/export conversions and also serves as the basis for the segmentation process yielding the "segmented transcription" format (EXS), which is used for search with the EXAKT tool.

The other two tiers appearing in the example (1) are **ref** (Reference) which stores an identifier for each sentence and **fe** (Free translation-English).

(1) Kamas

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 segmentation principles and the **ref** tier will be discussed in section 2.2; the principles of transcription and alternative transcription tiers (**st**, **stl**) in 2.3; the representation of multiple speakers and direct speech in 2.4. For the treatment of disfluencies and non-speech events see 4.2.

2.2. Segmentation levels: sentences, words and morphemes

2.2.1. Major units: sentences

It is becoming more and more common in current language documentation and discourse-oriented spoken language resources to recur to prosody-based segmentation units such as *intonation units* (intonation groups, intonation phrases) or *EDUs* (elementary discourse units) [Himmelmann 2006; Kibrik & Podlesskaya 2009; Izre'el & Mettouchi 2015], typically of a size of a clause or a simple sentence. Intonation units are considered descriptively much more adequate in representing spontaneous spoken discourse than grammatically and orthographically conceived "sentences" largely inherited from European written tradition. Experiments suggest that such segmentation might be produced reliably even by non-native speakers or non-speakers of the target language based solely on auditory cues [Himmelmann et al. 2018; Kibrik & Maisak 2020].

While fully recognizing the advantages of intonation units as "native" spoken language units, for practical reasons we could not adopt this approach in the INEL corpora. Instead, we use sentences as the main segmentation unit above the word level. The main reason behind this is the heterogeneity of the data comprised in each of the corpora.

Each INEL corpus so far has a share of audio recordings, indeed quite large for Dolgan and Kamas. A small part of the Dolgan recordings, though, represent literary prose read aloud rather than spontaneous discourse. The Kamas recordings of Klavdiya Plotnikova, in turn, document the last speaker, or rather a rememberer, of the language, and cannot be considered typical spoken data [Arkhipov & Däbritz 2018: 13–14]. The major part of Selkup and Evenki data, as well as the remainder of Dolgan and Kamas, are written materials — either field notes or edited folklore publications — for which no original sound is available in most cases. Dolgan, Kamas and Evenki written sources typically use orthography-based sentences.[8] The Selkup fieldnotes of Angelina Kuzmina do the same, while sometimes lacking any punctuation marks whatsoever. Therefore, we could not reconstruct any intonation-based transcription units for written sources, and for reasons of coherence we did not

[8] Some Evenki texts collected by Konstantin Rychkov seem to show two-level segmentation with major units (sentences) as combinations of smaller units (clauses). However, it only appears in part of the collection and cannot be extended to the rest of the data, organized more conventionally into sentences.

consistently attempt it for the remaining audio data either. Otherwise this would require two completely different systems of units within one corpus, making explode the complexity of both analysis and maintenance tasks.

The higher-level segmentation is thus based on orthographic sentences inasmuch as can be inferred from the written sources. For audio sources, the major unit is also a sentence, determined at the discretion of the researcher. In particular, it is not limited to a single predication or clause or rhythmic group, and long complex utterances such as (2) may be treated as one sentence, at the discretion of the researcher.

(2) Dolgan

ref	PoTY_2009_Aku_nar.038 (038)
ts	Ontuŋ momuja momuja voːpse baraːktiːr d͡iɛn h͡iɛse hil͡iɛbi, h͡iɛse totto bɨhɨlaːk, t͡uɔk toholaːk tot͡uɔj, de ol kačigirɨtar.
fe	*Then it [the reindeer] luckily goes away munching the bread, it probably had enough, how can it have enough? It crunches it.*

Every text in a corpus is thus represented as a sequence of sentences. Each sentence is assigned an identifier stored in **ref** tier. These IDs contain the text code, the speaker code (for multi-speaker texts only), and the consecutive number of the sentence. Additionally, the sentence number assigned by FLEx is also given in brackets. In many cases, the FLEx sentence number will be identical to the main sentence number ("038" in example (2) above). However, numbering in FLEx can include two levels (paragraph number.sentence number), as in (3). On the other hand, FLEx numbering in dialogues is consecutive irrespective of the speakers, whereas the main numbering will be independent for each speaker (4).

(3) Selkup

ref	SAlAn_1965_Soldatka_nar.**026** (**003.004**)
tx	Meː klupmɨn orsa som ɛːŋa.
fe	*Our club is very good.*

(4) Dolgan

ref-KuNS	PoPD_KuNS_2004_Life_conv.**KuNS.065** (001.208)
tx-KuNS	– Kanna atɨl͡iɨaktara onton karčɨ ɨl͡iɨaktara, eː?
fe-KuNS	– *Where do they sell, and they get money then?*

ref-PoPD	PoPD_KuNS_2004_Life_conv.**PoPD.144** (001.209)
tx-PoPD	– Mm, mm hiti kurdukkaːn anɨ.
fe-PoPD	– *Yes, yes, exactly like that now.*

The segmentation done by FLEx is normally preserved in the EXMARaLDA transcripts. However, it is sometimes altered manually to split or merge sentences. In this case, they are automatically renumbered to keep all IDs unique.[9] The original FLEx numbering (in brackets) is however preserved, corresponding to the first original FLEx sentence in case of merging.

The following annotation types are all aligned with sentences: alternative transcriptions beside the main transcription (section 2.3.3), translations and comments (section 3). These are all called sentence-level tiers below.

[9] The code for renumbering routine will be published together with other utilities under: https://gitlab.rrz.uni-hamburg.de/corpus-services/corpus-services.

2.2.2. Punctuation: orthography-based

At sentence level, the transcription is largely following orthographic conventions. It is designed so that the main transcription tier could be taken as a baseline for a (non-analytic) publication with only minor adaptations. We do not adhere to any detailed discourse transcription system.

The sentences are delimited by initial capitalization and sentence-final punctuation, not overlapping with the allowed sentence-internal punctuation. Direct speech is signaled by quotation marks, turn-taking in dialogues can be marked with a leading dash (see 2.3.3).

- The first word in a sentence is as a rule capitalized, as well as proper names, unless it appears problematic to find a matching capital letter for a given transcription symbol. Proper names are also capitalized on **mb** tier, while sentence-initial capitalization is not kept there.
- Allowed sentence-internal punctuation includes comma, dash (en-dash or em-dash, normally surrounded by spaces), semicolon, colon, quotation marks (either straight or matching): **, – — : ; " “ ”**. On **tx** tier, punctuation characters are included in the same event with the adjacent word. On all other word-level and morph-level tiers punctuation is omitted.
- Hyphen is allowed as word-forming character (see 2.2.3), although generally avoided.
- Single and double round brackets are reserved as special symbols (see 4.1 and 4.2); fragments of text inside brackets are treated specially.
- Allowed sentence-final punctuation includes full stop, question mark, exclamation mark and ellipsis: **. ? ! …** .[10] These can be combined with quotation marks but not between each other. Each sentence must end with a sentence-final punctuation. Note that ellipsis marks incomplete sentences; for marking of sentence-internal pauses and self-repairs see 4.2.1 and 4.2.2.

The segmentation algorithm implemented as a finite-state machine is published along with each corpus.[11]

2.2.3. Words and morphemes

Technically, the segmentation into meaningful units like words or utterances is independent from timeline events in the EXMARaLDA data model. However, in the INEL corpora word boundaries (and hence sentence boundaries) are forced to coincide with timeline events. This allows for easy matching between annotations across tiers.

Note that timeline events are strictly ordered but do not necessarily correspond to any absolute time values. Thus in texts with written source there are no time values at all, or they are all arbitrary. Texts with audio source are time-aligned at sentence level, while sentence-internal word boundaries are generally arbitrary.

More precisely, each event on **tx** tier includes a single word plus eventually any directly adjacent punctuation characters, and is ended with a space. As mentioned above, hyphen is allowed as word-forming character, although discouraged. It appears in cases where the orthography of the subject language or of the loanword source language would use a hyphen. This includes certain types of compounds and reduplication; combinations of a host word and a clitic; and proper names, especially borrowed ones:

- *bɨlɨr-bɨlɨr* “very long ago” [long.ago-long.ago], *ɨtɨ̈ː-ɨtɨ̈ː* “while crying” [crying-crying] (Dolgan)
- *gibər-n’ibud’* “somewhere” (Kamas; from *gibər* “where” + Russian *-нибудь*, indefinite clitic)
- *Sɨl’ča-Pɨl’ča* (reduplicative proper name, Selkup); *Köt’s’ün-Güd’ər* (proper name, Kamas)
- *Ust’-Az’örnɨj* (place name, Selkup, from Russian *Усть-Озёрное*)

[10] Ellipsis must be a single character, not three separate dots.
[11] See e.g. https://gitlab.rrz.uni-hamburg.de/inel-open-access/corpora/selkup/-/blob/main/corpus-utilities/segmentation.fsm

The segmentation of a word into morphemes is not handled by the EXMARaLDA segmentation algorithms. It is performed in FLEx during the glossing phase and imported with some adjustments. In EXB files, morpheme segmentation is present in the following tiers:

- **mb** (Morpheme breaks)
- **mp** (Morphophonemes (underlying))
- **ge** (Gloss-English), **gr** (Gloss-Russian), **gg** (Gloss-German, if present)
- **mc** (Morphological category)

It is stored within events corresponding to word boundaries on **tx** tier, explicit for human users but implicit for the EXMARaLDA software. Morpheme boundaries are signaled by one of the allowed delimiters — hyphen or, optionally for clitics, equals sign (**-** **=**). Zero morphs, which are assumed to have no overt exponent, are not represented in mb and mp tiers. Corresponding glosses and category labels in the **ge**, **gr**, **gg** and **mc** tiers are surrounded by square brackets and preceded with a leading dot instead of regular delimiters (**.[** **]**), and attached to the gloss of the preceding morph:

(5) Selkup

ref	PVD_1964_YoungBrother_nar.017 (001.017)				
tx	tɨmn'ä-u	kak	qaːr-ol'-dä	akoška-n	ɨl-o-ɣɨn
ge	brother**.[NOM]**-1SG	suddenly	shout-MOM-3SG.O	window-GEN	space.under-EP-LOC
fe	*Suddenly my younger brother burst out with a cry from under the window.*				

A consistent segmentation into morphemes across tiers is essential for the ISO/TEI export to function properly. Meanwhile, it can be accidentally broken while manually editing the corpus files. A dedicated routine is thus regularly run to check that the number of delimiters in each event matches across all morpheme-level tiers.[12] An INEL-specific version of the export into ISO/TEI format was created to explicitly express the segmentation into morphemes and to allow morpheme-based annotations to refer to each other across tiers [Ferger, Jettka 2020].

2.3. General transcription principles

The main transcription is governed by the following considerations.

2.3.1. Elementary units: Phonological/broad phonetic

At the word level, the transcription is aimed to be reasonably close to phonological. However, rarely can it be argued to be exactly phonological, for a number of reasons. There can be more or less elements of broad phonetic transcription, depending on the language.

- In some cases, the phonological analysis is not yet established, either for particular words or morphemes, or with respect to certain phonetic features of a language, such as the vowel length in Kamas. The Selkup materials contain a great deal of variation which can be ambiguous between dialectal variation, individual variation, and transcriber's variation. In such cases, some features are normalized in the main transcription based on our knowledge of particular dialects (e.g. stops do not have a voiced/unvoiced distinction in Northern Selkup), but other aspects may be kept unchanged from the original collector's transcription found in the manuscripts, unless we have a reason to do so.
- In some cases, variants which can be proven to be non-contrastive allophones are still transcribed differently if they are substantially different phonetically, perceived as distinct by

[12] The code for delimiter-checking routine will be published together with other utilities under: https://gitlab.rrz.uni-hamburg.de/corpus-services/corpus-services.

the speakers, and/or might be relevant for the analysis of particular phenomena such as the adaptation of Russian borrowings. As an example, the fricative **[ɣ]**[13] in Northern Selkup appears as an allophone of the uvular **/q/** and is represented as **q** in the main transcription. On the other hand, in Narym Selkup it can also stand for a realization of the intervocalic **/h/**, corresponding to **/s/** in other dialects. In order to enable further investigations of these sound relations, **ɣ** is preserved in the main transcription in Narym Selkup texts.

In cases of doubt, the original transcription stored in the **st** tier (see 2.3.3) and/or the provided sound recordings may be helpful to support a particular analysis.

2.3.2. Characters: Non-IPA

The character set is not that of International Phonetic Alphabet,[14] although a conversion into IPA would not be difficult to produce in an automated way. In particular, the following divergences from the IPA set are common:

- front rounded vowels are denoted with umlaut: **ä ö ü**
- the regular small **g** is used for the voiced velar stop instead of the "script g" **ɡ**
- palatalization is marked with an apostrophe **ʼ** (U+02BC Modifier letter apostrophe) rather than with superscript **ʲ**; palatals are also denoted with symbols for coronals plus apostrophe (e.g. **nʼ** instead of **ɲ**)
- hushing fricatives use caron; affricates are denoted with single characters rather than digraphs: **š ž č ǯ**

Note that in longer stretches of code-switching to Russian, Cyrillic Russian orthography is used instead of the Latin-based transcription (though some deviations from the standard pronunciation may still be reflected).

2.3.3. Alternative transcriptions

Apart from the two main transcription tiers, one or more additional transcription variants can be provided. For texts originating from written sources, a close rendition of the original is given in **st** (Source text) tier. For instance, in Selkup texts from Angelina Kuzmina's archive, it is the maximally exactly typed-in manuscripts (also provided in the corpus as scanned images of the corresponding pages). In addition to that, if the written source happens to be Cyrillic-based, then an automatically produced conversion into Latin characters is optionally provided in the **stl** (Source text-Latin) tier. While it might be identical to the one in **ts**/**tx**, the latter quite often contains more or less important manual corrections. Cf. an example in (6):

(6) Selkup

ref	KAI_1965_BoyAndOldDevil1_flk.003 (001.003)
st	ʻāрат йʻейты ʻшʼötты ʻныны kъ̊ʻннотын.
stl	**a**ːrat iːe**j**tɨ šötti nɨnɨ qənn**o**tɨn.
ts	**A**ːrat iːje**i**ːtɨ šötti nɨnɨ qənn**ɔ**ːtɨn.
fe	*In the autumn the sons went to the forest.*

For texts transcribed from the audio sources with the help of native speakers, their original transcription is provided in **stl** tier (if in Latin script), or in **st** tier (if in Cyrillic, in which case the optional **stl** tier

[13] This fricative is arguably more often uvular, thus IPA [ʁ]; however, Kuzmina only uses [ɣ] in her transcriptions.
[14] IPA Chart, http://www.internationalphoneticassociation.org/content/ipa-chart, available under a Creative Commons Attribution-Sharealike 3.0 Unported License. Copyright © 2018 International Phonetic Association.

may again contain an automatic conversion into Latin). The main transcription in **ts**/**tx** tiers then contains further corrections by the researchers (cf. (7), (8)).

(7) Selkup

ref	YIF_196X_WorkAsPostwoman_nar.007 (007)
st	Нашақоуттэ нетуа ватт, а только аштэҳе қваяқӷуҳут.
stl	Našaq**ou**tte **n**etua watt, a tol'ko aš**te**x**x**e qwaja**qq**u**x**ut.
ts	Našaq**o**tte **n**'ej**t**u**h**a watt**ə**, a tol'ko aš**te**h**h**e qwaja**kk**u**h**ut.
fe	*There was no road then, we would ride the reindeer.*

(8) Dolgan

ref	MiPP_1996_OldManButterfly_flk.015 (001.015)
st	Дьэ, аһааннар, **дьэ** онтуларын астара баранна – Лөрүө гиэнэ.
ts	D'e, aha:nnar, ontularɨn astara baranna, Lörüö g͡iene.
fe	*Well, having eaten, their food ran out, Butterfly's one.*

2.4. Speakers, turn-taking and direct speech

A larger part of INEL data is monological, but each corpus so far has at least some dialogues, most of which are radio interviews in the Dolgan corpus. In principle, each speaker in dialogues is transcribed and annotated in its own set of tiers, so as to naturally represent possible overlapping of speech. The speaker is then marked with the corresponding speaker code in tier names. The speaker code(s) of the principal speaker(s) is always part of the text name. (Codes starting with "NN" are reserved for unknown speakers.)

(9) Selkup

ref-NP	**YIF_NP**_196X_WeShouldMakeBaskets_conv.**NP**.025 (050)
ts-NP	Toʒe qöškəlaj?
fe-NP	*Shall we go there, too?*

ref-YIF	**YIF_NP**_196X_WeShouldMakeBaskets_conv.**YIF**.026 (051)
ts-YIF	Mi tanoptə wez'd'e qwajakkulajhe patom.
fe-YIF	*You and me, we'll go everywhere later.*

However, if only one speaker is the principal narrator while some other voice(s) appear occasionally, such as researcher's prompts or listener's remarks, they may be transcribed in the same set of tiers. In such cases, a change of speaker is marked with speaker code followed by a colon and surrounded with double brackets, e.g. as in (10). In translation tiers, square brackets are used instead of double round brackets.

(10) Kamas

ref	PKZ_196X_FrogPrincess_flk.012 (014)	PKZ_196X_FrogPrincess_flk.013 (014)
ts	**((PKZ:))** Saʔməluʔpi…	**((KA:))** Боярский.
fe	***[PKZ:]*** *It fell down…*	***[KA:]*** *A boyar's [palace].*

ref	PKZ_196X_FrogPrincess_flk.014 (014)
ts	**((PKZ:))** băjarskəj dvărestə.
fe	***[PKZ:]*** *…near a boyar's palace.*

Direct speech is surrounded by quotation marks (either straight " " or matching “ ”).

(11) Kamas

ref	PKZ_196X_Alenushka_flk.005 (006)
ts	Dĭ n’i măndə: "**Măn ugaːndə bü bĭssittə axota**".
fe	The boy says: "I want very much to drink water [= I am very thirsty]."

Turn-taking in dialogues can be signaled by an en-dash or em-dash (– —). While it is also explicitly represented by tier separation, it is generally optional. However explicit marking is more helpful in reported dialogues; alternating quotes maybe used in this case.

(12) Kamas

ref	PKZ_196X_TopsAndRoots_flk.001 (001)	PKZ_196X_TopsAndRoots_flk.002 (002)
ts	On’iʔ kuza kuʔpi r’epa.	A urgaːba šobi.
fe	*One man sowed turnip.*	*And a bear came.*

ref	PKZ_196X_TopsAndRoots_flk.003 (003)	PKZ_196X_TopsAndRoots_flk.004 (004)
ts	"Ĭmbi alaʔbəl?"	"Kuʔlaʔbəm r’epa".
fe	*"What are you doing?"*	*"I am sowing turnip."*

3. Translations and comments: general principles

3.1. Main translation and comments tiers

The INEL project is based in Germany, the working language of the project and the international language of linguistics being English, and most of the linguistic resources in scope of the project using Russian to some extent. Thus English, German and Russian are the three analysis or annotation languages used in the INEL corpora, with preference to English when a choice is to be made. They are used in sentence-level translation tiers, in comments as well as in lexical glosses in glossing tiers.

The main INEL translations found in tiers (resp. **fe**, **fg**, **fr**) aim to be (i) grammatically correct and idiomatic in the target language, while (ii) keeping reasonably close to the original structure. Generally, translation into each of the languages is intended to be readable on its own (without considering other translation tiers or glosses) as a self-contained text or story. It is however not guaranteed to reflect the syntax, morphological categories (such as tense) or lexical patterns of the original.

Detailed information on the original linguistic form can partially be inferred from the glosses but should ideally be complemented by lengthier grammatical descriptions of the subject language. Specifically problematic cases are addressed in comments (tier **nt** (Notes)). The comments are only written in English and are not duplicated in other languages. Each comment is preceded with the code of its author in square brackets:

(13) Selkup

ref	KFN_196X_DrunkBear2_nar.006 (006)	
tx	qorqə	šer-ba-t
ge	bear.[NOM]	get.drunk-PST.NAR-3PL
fe	*They made the bear drunk.*	
nt	[WNB:] Why is the verb in the objective conjugation? [AAV:] šerbat: palat. r^j	

Comments related to the entire text are generally found with its first sentence. These may contain references to related texts in the corpus, references to the original text in case of translations or retellings, etc.

3.2. Alternative translations and comments

When available, the translations provided by the original collectors or publishers of the material, or by native speakers during transcribing sessions, are given in supplementary tiers **ltr** and **ltg** resp. for Russian (e.g. in Selkup and Dolgan corpora) and German (e.g. in the Kamas corpus).

The comments by original collectors or publishers, if present, are given in the tier **nto** (Notes-Original).

4. Transcription and translations: special cases

The system reflected below is driven by the desire to keep special characters and patterns to a minimum, preserving to the extent possible the "normal" orthographic layout of transcriptions and translations. The aim was to lessen the burden for the reader, making the text also suitable for eventual extraction and publication without the rest of the interlinear annotations. On the other hand, we tried to capture the most prominent phenomena intrinsic (i) to the oral discourse in our audio sources, (ii) to manuscript field notes in our written sources, and (iii) to the parallel texts formed by transcriptions and translations together.

The main special characters are single and double round brackets. Punctuation marks inside brackets is not considered for segmentation into sentences, in particular question marks and ellipsis characters are not treated as sentence-final.

4.1. Uncertainty in form and meaning

4.1.1. Uncertainty

Round brackets (parentheses) are suggested as the main symbol to mark uncertainty in transcription, which is possible to the extent that they are not expected to be encountered in their common orthographic function in the INEL corpora. It is supplemented by question marks in translations.

Uncertain forms

A fragment which is transcribed but with uncertainty is surrounded by single round brackets in transcription: **( )**. The associated uncertainty in the sentence-level translation tier is marked with brackets and a question mark: **(?)**.

(14) Constructed example

tx	Qarra ittiti (nimti).
fe	*He brought it (there?) to the river.*

Certain forms with unknown/uncertain meaning

When a word is reliably transcribed but unknown or seemingly irregular, it is not specifically marked in transcription.

On morph-level tiers, unknown elements are marked with **%%** (esp. in glosses). Tentative glosses suggested for unknown elements are marked by a single leading percent sign, e.g. **%snake**.

On sentence level, the missing translation for an unknown element is marked with a question mark in round brackets: (?). Suggested but uncertain translations are surrounded with brackets and a question mark, as in the preceding case.

(15) Constructed example

tx	Ma:ša	čuren'n'e	i	qa:ǯiq	kuranna	n'a:blaštjatko.
ge	Masha	cry.FRQ.AOR	and	%%	run.AOR	duck.DU.TRL
fe	*Masha burst into tears, and ran (?) after the ducks.*					

(16) Constructed example

tx	Ma:šen'ka	čaɣɨlǯid	pöp	a:rɣond.
ge	Mashenka	%%-INT.PF.3SG	stone.ACC	other.ILL
fe	*Mashenka (moved?) the stone aside.*			

Alternative transcriptions or translations

Two or more equally uncertain options are surrounded by single brackets and separated by a slash, without spaces:

(17) Constructed example

tx	Ka (mompa/montɨ) mɨ tökam isamɨ.
fe	*Let me take the goose, he said.*

If the alternatives call for different translations, the translations should also be separated with a slash:

(18) Constructed example

tx	Ka mompa mɨ (tökam/č'iŋkɨm) isamɨ.
fe	*Let me take the (goose/swan?), he said.*

A secondary alternative to the main transcription variant is surrounded by single round brackets with a slash after the opening bracket. This option is mostly useful for data from manuscripts where some alternate forms are written above or beside the main transcription line.

(19) Constructed example

tx	Qumɨt kosti (/mɔ:ttɨla) tüntɔ:tɨt.
fe	*People come to visit us.*

If the alternatives call for different translations, the markup in translation mirrors the markup in transcription:

(20) Constructed example

tx	Ka mompa mɨ tökam (/č'iŋkɨm) isamɨ.
fe	*Let me take the goose (/swan), he said.*

4.1.2. Unintelligible fragments

Completely unintelligible fragments are marked with ellipsis character surrounded with double round brackets.[15] In translation, the missing fragment is marked with ellipsis in single round brackets.

(21) Constructed example

tx	Č'unt̶i̶ č'ot̶i̶t ((…)) mittam.
fe	*I will give (…) for the horse.*

Partially unintelligible fragments can be rendered with the recoverable portion transcribed as usual, with ellipsis to indicate incompleteness, surrounded by double round brackets. In translation, ellipsis in single round brackets is used, same as in the previous case.

(22) Constructed example

tx	Č'unt̶i̶ č'ot̶i̶t ((so…)) mittam.
fe	*I will give (…) for the horse.*

4.1.3. Unclear sentences

Sentences whose meaning is entirely unclear bear a question mark in square brackets **[?]** in the end of or instead of translation:

(23) Kamas

ref	NN_1914_Birds_flk.003 (001.004)
tx	Măn toru inen toru kunda tora to?la? kunnim."
fe	*I will beat brown on the brown mane of a brown horse.* ***[?]***
fg	*Ich werde die braune Mähne eines braunen Pferdes braun prügeln.“* ***[?]***

4.2. Disfluencies and non-speech events

Since INEL corpora do not focus on the study of speech production, we only reflect and distinguish a very limited range of phenomena in our transcriptions. However, at least some degree of accuracy is desired to allow users to better match the transcription line with the sound, and also to eventually enable further automated processing (e.g. omit all disfluencies for visualization, for syntactic analysis, etc.).

An ideal transcription system would be easy to type and read, while capturing all kinds of the selected phenomena. This already yields some non-trivial requirements. On top of that, some additional limitations come from the software which is used in the project. Namely, both FLEx used for glossing and EXMARaLDA used for further annotating and corpus search impose their own limitations on possible sequences of characters allowed in transcription.

4.2.1. Non-speech sounds and events

Non-speech sounds such as laughter, coughing, taking breath etc., as well as long pauses, or extraneous noise, tape breaks etc. are marked with short mnemonic labels in capital letters surrounded by double brackets, e.g.:

[15] Ellipsis must be a single character, not three separate dots.

(24) Constructed example

tx	Qumɨt kosti ((COUGH)) tüntɔːtɨt.
fe	*People come to visit us.*

The list of labels is open and includes at least the following:

- BREATH
- COUGH
- LAUGH
- PAUSE
- NOISE
- DMG (signals damaged sound recording)
- BRK (signals a break in recording, i. e. when the recording was stopped and restarted)

4.2.2. False starts/self-repairs

False starts (words rejected by self-repair, *Reparatur*) are marked with a single trailing dash indicating incompleteness and enclosed in single round brackets.

(25) Constructed example

tx	Qumɨt (ko-) kosti tüntɔːtɨt.
fe	*People come to visit us.*

Truncated words have no interlinear glossing. However, complete words or identifiable stems can be glossed (see (27) below). If a word rejected by the speaker is complete (not truncated), it is marked with a trailing equals sign and likewise enclosed in single round brackets.

(26) Constructed example

tx	Qumɨt (kosti=) kosti tüntɔːtɨt.
fe	*People come to visit us.*

Multiple consecutive false starts can be enclosed in one pair of brackets.

(27) Kamas

ref	PKZ_196X_FireBird_flk.088 (090)						
ts	Da d'abit inem, a uzdam **(ej= uzu- i?= užum- i? u-)** i? i?!						
tx	Da	d'abi-t	ine-m,	a	uzda-m	**(ej=**	**uzu-**
ge	and	capture-IMP.2SG.O	horse-ACC	and	halter-NOM/GEN/ACC.1SG	NEG	

tx	**i-?=**	**užum-**	**i-?**	**u-)**	i-?	i-?
ge	NEG.AUX-IMP.2SG		NEG.AUX-IMP.2SG		NEG.AUX-IMP.2SG	take-CNG
fe	*And catch the horse, but don't take the halter!*					

4.2.3. Annotator corrections

When a form which is pronounced clearly is judged by the transcriber/annotator as inappropriate (non-standard, ill-formed, grammatically or lexically out of context, etc.), it is not corrected in the transcription but reported in the comments instead. Glosses may also be supplied in the comment line if needed, following the suggested correction in square brackets:

(28) Dolgan

ref	ErTS_AkPG_1994_AAPopov_nar.ErTS.003 (001.007)
tx	Bi:r kinige komullubataktara kahan da.
fe	*They were never collected in one book.*
nt	[DCh]: "kinige" is probably a haplology of "kinige-ge" [book-DAT/LOC], which would be the expected form.

4.3. Mismatches between original text and translation

4.3.1. Reconstructed referents

Referents unspecified or underspecified in the source language but essential for the correct interpretation are explicated in square brackets.

(29) Kamas

ref	AA_1914_Brothers_flk.004 (001.004)			
tx	So	abii?,	sonə	sarbii?.
ge	raft.[NOM.SG]	make-PST-3PL	raft-LAT	bind-PST-3PL
fe	***[The people]*** *made a raft, tied them* ***[the children]*** *to the raft.*			

In the above example, the sentence remains unclear from the context, since the two participants (the people and the children) are only signalled by 3 pers. plural verbal agreement, while for *the people* it is their first mention in the text.

4.3.2. Material added or omitted for grammaticality

Words added or omitted to preserve grammaticality in the target language, especially proforms and adverbs, are left unmarked.

(30) Kamas

ref	AA_1914_Corpse_flk.071 (003.036)			AA_1914_Corpse_flk.072 (003.037)		
tx	"No,	i?	t'ora?!	Bar	mĭlim	sumnam."
ge	well	NEG.AUX-IMP.2SG	cry-EP-CNG	all	give-FUT-1SG	five-ACC
fe	*"Come on, don't cry!*			***I*** *will give* ***you*** *all five."*		

A reverse case is removing an element which would be superfluous in the translation language (cf. possessive marking in the address form):

(31) Kamas

ref	AA_1914_Corpse_flk.011	AA_1914_Corpse_flk.012 (002.007)			
tx	"Ad'am!	Aspa? edə?, uja pada?!"			
ge	aunt-NOM/GEN/ACC.1SG	cauldron.[NOM.SG]	hang.up-IMP.2SG	meat.[NOM.SG]	cook-IMP.2SG
fg	„Tante!	Häng einen Kessel auf und koche Fleisch.“			
ltg	*»****Meine*** *Tante!*	*Einen Kessel hänge (über das Feuer), Fleisch stecke (hinein)!»*			

4.3.3. Significant material added for clarification

Information missing in the source language but important for interpretation can be added in square brackets. If it is only suggested but not reliable, a question mark is added before the closing bracket.

(32) Dolgan

ref	ErSV_1964_WarBirdsAnimals_flk.520			
tx	Bajgaltan	min	ɨla	kiːri͡ektere.
ge	river-ABL	soup.[NOM]	take-CVB.SIM	go.in-FUT-3PL
fe	*They will go to the river to take water **[for the soup]**.*			

(33) Kamas

ref	AA_1914_Corpse_flk.064 (003.029)			
tx	Bazoʔ	tʼorlaʔ	tĭrlöleʔ	kujobi.
ge	again	cry-CVB	roll-CVB	stay-PST.[3SG]
fe	*Again crying he starts rolling **[into the fire]**.*			

4.3.4. Literal vs. idiomatic translation

If the intended meaning of the sentence is hard to infer from the lexical meanings of the words, both may be combined in one translation, first the literal meaning followed by the intended (e.g. idiomatic) meaning given in square brackets with a leading equals sign (**[=]**):

(34) Kamas

ref	PKZ_196X_FoxAndHare_flk.005 (005)			
tx	Lisan	turat	bar	mʼaŋŋuʔpi.
ge	fox-GEN	house-NOM/GEN.3SG	PTCL	flow-MOM-PST.[3SG]
fe	*The fox's house **flowed away [=melted]**.*			

5. Annotation of borrowings, calques and code-switching

The annotation of borrowings and code-switching, also present in the Nganasan corpus [NSLC], has been redesigned for the INEL corpora and will be outlined in the following sections.

First of all, a disclaimer should be made about the separation of these phenomena in the annotation. It is well known that strict differentiation between borrowings and code switching is not possible, and probably even not always desirable. Thus there are dubious cases which can be interpreted differently; sometimes, but by no means always, new data can be helpful to decide on the status of a particular element, e.g. if it reveals to be regularly used by other speakers.

We have thus adopted simple ad-hoc annotation rules, keeping in mind that particular decisions are subject to change a posteriori.

- Long stretches of speech in another language (mostly Russian) are naturally classified as code-switching.
- Single words that demonstrate phonological and morphological adaptation, are inflected according to the matrix language rules, and/or tend to occur often, are naturally classified as borrowings.
- Multi-word units such as proper names (e.g. Angelina Ivanovna) or collocations (e.g. *sovetskij vlast'* "Soviet power") are treated as borrowings, each word is annotated separately.
- Single words which appear infrequently and do not show evidence of adaptation, i.e. the "grey zone", are by default treated as borrowings. However they can be simultaneously annotated as code-switching, which makes it easier to reassess their status later.

Thus the **CS** (Code-switching) tier will be primarily marked for sequences of more than one word, or separate one-word clauses or quotations. Types distinguished in CS tier are dealt with in section 5.2, while section 5.1 considers borrowings.

Yet another related phenomenon, namely calques, that is contact-induced grammatical patterns without borrowing of lexical material, are annotated in the same tier as code-switching and are treated in section 5.2.2.

5.1. Borrowings

Borrowings are annotated on up to three tiers. The annotated features include:

- source language (**BOR**)
- semantic/lexical class (**BOR**)
- phonological adaptations (**BOR-Phon**)
- morphological adaptations and inflection (**BOR-Morph**)

To reduce the number of annotation tiers, source language and semantic/lexical class annotations are combined within the primary **BOR** tier. Since both kinds of information involve only lexical information and require no context knowledge, they are commonly introduced in the FLEx lexicon and exported along with the glossing. The other two tiers are optional and can only be annotated based on the specific wordform shape.

5.1.1. Source language

The source language of borrowing, or a language group if the specific language is unknown, is designated by an abbreviated language name followed by a colon. The part after the colon encodes semantic/lexical type (see next section).

The following labels are currently in use:

- **RUS:** Russian
- **TURK:** unspecified Turkic language
- **DOLG:** Dolgan (Turkic)
- **KHAK:** Khakas (Turkic)
- **TAT:** Tatar (Turkic)
- **YAK:** Yakut (Turkic)
- **DOLG/YAK:** Dolgan or Yakut
- **NEN:** Nenets (Uralic)
- **SELK:** Selkup (Uralic)
- **EV:** Evenki (Tungusic)
- **KET:** Ket (Yenisseian)
- **MONG:** unspecified Mongolic language

In case of doubt, a suggested source language abbreviation is preceded by percent sign, e.g. **%TURK**.

5.1.2. Semantic/lexical class of borrowing

The NSLC annotation scheme [Wagner-Nagy et al. 2018] distinguishes between just two semantic types of borrowings, viz. cultural and core borrowings. The INEL corpora extends these two classes, relevant mainly for content/lexical borrowings, with a few narrower classes pertaining more to function words.

- **cult** cultural borrowing
- **core** core borrowing
- **gram** grammatical device
- **mod** modal word
- **disc** discourse marker

Most borrowed nouns and verbs are labelled **cultural** borrowings. These naturally include names of artifacts, proper names, names of professions/occupations (e.g. ‘doctor’, ‘teacher’), as well as names of plants and animals (except generic terms such as ‘fish’ or ‘tree’), types of activities introduced with the

contacting/dominant culture (such as ‘to study (at school)’, ‘to vote’). This is the default and the most numerous borrowing class.

Rare **core** borrowings concern very basic vocabulary, for which native alternatives also exist. These include kinship terms, body part names, basic qualities (e.g. ‘good’, ‘young’). We also include here universal quantifiers (e.g. ‘all’, ‘every’) and numerals.

As an example of an interesting borderline case, for Selkup we annotate the borrowed Russian form for ‘year’ as in “in year 1954” as a cultural borrowing, since the Selkup, as well as some other Samoyedic peoples, seem to have lacked the concept of a calendar year (as opposed to months/seasons).

(35) Selkup

ref	TVP_1965_IWasBornInChaselka_nar.002 (001.002)					
tx	1954	godu	qessak	N’aral’	Mač’ont	uč’itɨqa.
ge	1954	in.year	leave-PST-1SG.S	tundra-ADJZ	forest-ILL	study-INF
BOR	RUS:cult	**RUS:cult**				RUS:cult
fe	*In 1954 I went to school in Krasnoselkup.*					

Three further subtypes are grossly labelled **grammatical** devices (e.g. conjunctions “and”, “if”), **modal** words (e.g. “maybe”, “should”) and **discourse** markers (e.g. particles like Russian *ну* “well”). It must be stressed that this threefold distinction is to a large extent arbitrary and by no means clear-cut; it is taken merely as a starting point for observations and eventual modifications.

Several subgroups can be distinguished within these three classes, some of them specifically treated in [Matras 2011]. **Discourse** markers are tightly connected to speaker interaction in dialogue; they are placed quite high on the borrowability scale.

- speech formulas: “hello”, “thank you”; “yes”, “no”
- interjections: “Oh my God!”, “Quiet!”
- various particles, often hardly translateable: e.g. Russian *ну* “well”, *же* (emphatic), *ведь* “after all”

The class of **modal** words is perhaps the most heterogeneous in the present classification. It includes several groups which may behave differently with respect to borrowability, notably:

- modal words: Russian *нужно* “(one) should”, *надо* “(one) needs”, *охота* “(one) wants”, *может* “maybe”
- focus particles: “even”, “too”, “only”
- phase adverbs: “already”, “still”

(36) Dolgan

ref	AkEE_19XX_BoySister_flk.137 (001.129)			
tx	"Min	ed’i:jbin	bulu͡okpun	na:da.
ge	1SG.[NOM]	older.sister-1SG-ACC	find-PTCP.FUT-1SG-ACC	one.should
BOR				RUS:mod
fe	*"I have to find my sister.*			

The class of **grammatical** devices is also not uniform. The group of conjunctions is particularly prominent; e.g. in the whole Kamas corpus, the three coordinating conjunctions *и, да* “and”, *а* “and, but” constitute approximately 30% of the total occurrences of borrowings (ca. 2400 out of 8300). Other types of conjunctions and connectors are significantly less prominent in terms of occurrences, although still well represented. Borrowed derivational affixes are noted in the INEL Dolgan corpus (diminutive) and marginally in Evenki, and inflectional affixes (Turkic plural *-LAr* and possessive *-Im*) in Evenki.

In case of affixes and clitics, the relevant gloss is given in round brackets after the usual borrowing tag, in order to disambiguate with borrowed stem marking.

(37) Dolgan

ref	BeAM_199X_LegendSpiritOfTrees_nar.154 (001.151)					
tx	**Hopka-kaːn**	baːr	e-t-e	hugas-kaːn	d'i͡e	atti-gar
ge	**hill-DIM**.[NOM]	there.is	be-PST1-3SG	close-INTNS.[NOM]	house.[NOM]	place.beneath-DAT/LOC
BOR	**RUS:core** **EV:gram(DIM)**					
fe	*There was a small hill very close, beneath the house.*					

To sum up, the following groups of borrowed items are labelled as grammatical:

- coordinating conjunctions: "and", "but", "or"
- other conjunctions and connectors: "if", "that", "in order to", "otherwise"
- adpositions: "across", "until"
- grammaticalized periphrastic markers (mood, phase): Russian *давай* (hortative), *пускай* (jussive), *бы* (irrealis); *давай* (inchoative), *бросить* (cessative)
- indefinite and negative pronoun series markers: "some", "any", "no"
- derivational affixes: diminutive, intensifier

Grammatical borrowings in this sense are to be distinguished from grammatical elements (e.g. number or case markers) appearing on borrowed/embedded items.

Among the possible adjustments of the described system, one can suggest considering a group of universal quantifiers, indefinites and negatives, currently divided across the **core** and **grammatical** classes. This includes expressions like "each", "all", "every", "always" and indefinite/negative pronouns or indefinite/negative series markers. Borrowings of this group are represented in all the current INEL corpora. Morphological markers (borrowed affixes) might also be considered as a separate class.

5.1.3. Phonological adaptation

Phonological adaptations of borrowings can involve one or more of the processes listed below:

BOR-Phon	**Types**	**Annotation tag**	**Comment**
Adaptation strategies	deletion	**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)
	insertion	**inVins**	initial vowel insertion
		medVins	medial vowel insertion
		finVins	final vowel insertion
	substitution	**Csub**	consonant substitution
		Vsub	vowel substitution
	lenition	**lenition**	weakening
	fortition	**fortition**	strengthening

5.1.4. Morphological adaptation

The two parameters encoded here are (i) morphological adaptation strategy needed to make a foreign lexeme function as a native word, and (ii) inflection in matrix language, i.e. whether the item from the

embedded language is further inflected in matrix language or appears as bare (perhaps adapted) stem. They are encoded by respective tags separated with a colon.

BOR-Morph	Types	Annotation tag	comment
Adaptation strategies	direct insertion	**dir:**	insertion without any morphological adaptation
	indirect insertion	**indir:**	insertion with morphological adaptation
	paradigm insertion	**parad:**	a (subset of) paradigm of inflected forms is borrowed, e.g. SG vs PL forms, PRS vs PST forms
Further inflection	bare	**:bare**	item can be adapted but not inflected in matrix language
	inflected	**:infl**	item is further inflected in matrix language

5.2. Code-switching and calques

5.2.1. Code-switching

The annotation tags used for code-switching in **CS** tier start with the abbreviated name of the language followed by a colon, similar to **BOR** tier. Note that the CS tier is also filled for researchers or other participants in the recording who do not speak the target language and thus do not "switch", technically speaking. The following language abbreviations are currently in use:

- **RUS:** Russian
- **DOLG:** Dolgan
- **EST:** Estonian
- **FIN:** Finnish
- **KHAK:** Khakas
- **EV:** Evenki

The second part of the code-switching tag encodes one of the structural types (or, alternatively, signals a calque; see 5.2.2).

Types	Annotation tag	Comment
sentence-external CS	**:ext**	languages switch at sentence (clause, utterance) borders and do not interfere; also valid for quotations (direct speech)
sentence-internal CS	**:int.ins**	insertion: a fragment in embedded language is inserted replacing a part of structure in matrix language – e.g. a noun phrase, a time adjunct, etc.; languages change at phrase borders
	:int.alt	alternation: two languages switch at an arbitrary point, the fragment in embedded language does not form a syntactic unit
	:int	a single word is inserted, distinguishing between subtypes is problematic

5.2.2. Calques

Constructions which are atypical for the main language and are likely to be modelled after another language (here Russian) are marked as calques (loan translations). Code-switching and calques are annotated in the same tier, **CS**, since they are not expected to occur simultaneously. E.g. if there is code-switching to Russian, no other features will be annotated in this tier except code-switching type.

Vice versa, if the construction in the main language is judged to be a calque from Russian, it is uttered in the main language and not in Russian (hence no code-switching is marked).

No types are distinguished for calques, thus the annotation scheme only includes source language (typically Russian) and the "**:calq**" tag.

(38) Kamas

ref	PKZ_196X_MiceBuryCat_flk.003 (003)				
tx	Dĭgəttə	dĭ	ibi	da	küla:mbi.
ge	then	this.[NOM.SG]	take-PST.[3SG]	and	die-RES-PST.[3SG]
BOR				RUS:gram	
CS			**RUS:calq**		
fr	*Потом он взял да и умер.*				
fe	*Then it suddenly died.*				
nt	[GVY:] a calque of the Russian "взял и…" 'suddenly…'.				

References

Arkhangelskiy, T. & Ferger, A. & Hedeland, H. 2019: Uralic multimedia corpora: ISO/TEI corpus data in the project INEL. In: *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*. P. 115–124.

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

Arkhipov, Alexandre & Däbritz, Chris Lasse & Gusev, Valentin. 2020: User's Guide to INEL Kamas Corpus. *Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology*, 3, 1–34. DOI: https://doi.org/10.14232/wpcl.2020.3.

Brykina, Maria; Orlova, Svetlana; Wagner-Nagy, Beáta. 2020. INEL Selkup Corpus. Version 1.0. Publication date 2020-06-30. Archived in Hamburger Zentrum für Sprachkorpora. http://hdl.handle.net/11022/0000-0007-E1D5-A. In: Wagner-Nagy, Beáta; Arkhipov, Alexandre; Ferger, Anne; Jettka, Daniel; Lehmberg, Timm (eds.). *The INEL corpora of indigenous Northern Eurasian languages*.

Däbritz, Chris Lasse. 2020: User's Guide to INEL Dolgan Corpus. *Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology*, 4, 1–52. DOI: https://doi.org/10.14232/wpcl.2020.4.

Däbritz, Chris Lasse; Kudryakova, Nina; Stapert, Eugénie. 2019. INEL Dolgan Corpus. Version 1.0. Publication date 2019-08-31. http://hdl.handle.net/11022/0000-0007-CAE7-1. 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*.

Dickinson M.; Tufiş D. 2017. Iterative Enhancement. In: Ide N., Pustejovsky J. (eds). Handbook of Linguistic Annotation. Springer, Dordrecht. https://doi.org/10.1007/978-94-024-0881-2_9.

Ferger, Anne & Jettka, Daniel. 2020. Use Cases of the ISO Standard for Transcription of Spoken Language in the Project INEL. In: *Proceedings of CLARIN Annual Conference 2020*. Eds. C. Navarretta and M. Eskevich. Virtual Edition. P. 126–130.

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*.

Himmelmann, Nikolaus P. 2006: The challenges of segmenting spoken language. In: Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*. Berlin: Mouton de Gruyter. P. 253–274.

Himmelmann, Nikolaus P. & Sandler, Meytal & Strunk, Jan & Unterladstetter, Volker. 2018: On the universality of intonational phrases: A cross-linguistic interrater study. *Phonology*, 35(2). P. 207–245. https://doi.org/10.1017/s0952675718000039.

Izre'el, Shlomo & Mettouchi, Amina. 2015: Representation of speech in CorpAfroAs: Transcriptional strategies and prosodic units. In: *Corpus-based Studies of Lesser-described Languages: The CorpAfroAs corpus of spoken AfroAsiatic languages*. Amsterdam: John Benjamins. P. 13–41. [Available online at https://www.tau.ac.il/~izreel/publications/IZRE'EL&METTOUCHI REPRESENTATION OF SPEECH.pdf]

Kibrik, A. A., & Podlesskaja, V. I. (eds.) 2009: *Rasskazy o snovidenijax: Korpusnoe issledovanie ustnogo russkogo diskursa* [Night Dream Stories: A corpus study of spoken Russian discourse]. Moscow: Jazyki slavjanskix kul'tur.

Кибрик, А. А. & Майсак, Т. А. 2020: Московские правила дискурсивной транскрипции для описательных и документационных исследований. In: *Звегинцевские чтения — 2020: К 60-летию кафедры и отделения теоретической и прикладной лингвистики и 110-летию со дня рождения В. А. Звегинцева. Материалы конференции (Москва, МГУ имени М. В. Ломоносова, 30–31 октября 2020 г.)*. Moscow. P. 58–60.

Matras, Yaron. 2011. Universals of structural borrowing. In: Siemund, Peter (ed.). Linguistic universals and language variation. Berlin: Mouton. P. 200–229.

Mettouchi, Amina, & Chanard, Christian. 2010: From Fieldwork to Annotated Corpora: The CorpAfroAs project. *Faits de langues*, (2).

Schmidt, Thomas & Wörner, Kai. 2014: EXMARaLDA. In: *Handbook on Corpus Phonology*. Oxford University Press. P. 402–419.

Sloetjes, H. & Wittenburg, P. 2008: Annotation by category — ELAN and ISO DCR. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

Wagner-Nagy, Beáta & Szeverényi, Sándor & Gusev, Valentin. 2018: User's Guide to Nganasan Spoken Language Corpus. *Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology*, 1 (1), 1–45. DOI: https://doi.org/10.14232/wpcl.2018.1.

Appendix A. Complete tier layout examples

(1) Monologue (Selkup)

	264	265	266	267	268	269	270	271	272	273
ref	KAI_1965_OldManWithLittleMind1_flk.049 (002.021)									
st	**ʹкоты яйцаты то много, чинкыш ʹйса҆мёна (ʹёԓтоко) [ʹеԓты чоты].**									
stl	**qoti ɨjcati to mnogo, činkɨp iːsamena (eŋtoqo) [eŋti čoti].**									
ts	Qɔːti jajcati to mnogo, čʼɪŋkɨp iːsam ɛna (ɛŋtoːqo) (/ɛŋti čʼɔːti).									
tx	Qɔːti	jajcati	to	mnogo,	čʼɪŋkɨp	iːsam	ɛna	(ɛŋtoːqo)	(/ɛŋti	čʼɔːti).
mb	qɔːti	jajca-ti	to	mnogo	čʼɪŋkɨ-p	iː-sa-m	ɛna	ɛŋ-toːqo	ɛŋ-ti	čʼɔːti
mp	qɔːti	jajca-ti	to	mnogo	čʼɪŋkɨ-m	iː-sɨ-m	ɛnä	ɛŋ-ntoːqo	ɛŋ-ti	čʼɔːti
ge	probably	egg.[NOM]-3SG	EMPH2	many	swan-ACC	take-PST-1SG.O	CONJ	egg-TRL.3SG	egg.[NOM]-3SG	for
gr	наверно	яйцо.[NOM]-3SG	EMPH2	много	лебедь-ACC	взять-PST-1SG.O	CONJ	яйцо-TRL.3SG	яйцо.[NOM]-3SG	для
mc	ptcl	n-n:case-n:poss	ptcl	quant	n-n:case	v-v:tense-v:pn	ptcl	n-n:case.poss	n-n:case-n:poss	pp
ps	ptcl	n	ptcl	quant	n	v	ptcl	n	n	pp
SeR		np:Th			np:Th	0.3.h:A				
SyF		np:S		n:pred	np:O	0.3.h:S v:pred				
IST		accs-Q			giv-active-Q			giv-active-Q		
BOR		RUS:core		RUS:core						
BOR-Phon										
BOR-Morph		parad:infl		dir:bare						
CS			RUS:int.alt							
fr	"Наверное яиц-то много, лебедя я бы взял (ради яиц)".									
fe	"I could probably get a lot of eggs, I would take a swan for the eggs".									
fg	"Ich könnte wohl viele Eier bekommen, ich würde einen Schwan nehmen wegen der Eier."									
ltr	наверное лебедя взял бы за яйца									
nt										
nto										

(2) Dialogue (Dolgan)

	38 [00:24.4]	39 [00:24.9]	40 [00:25.4]	41 [00:2	42 [00:26.4]
ref-KuNS	ElBK_KuNS_2004_Storytellers_conv.KuNS.004 (001.006)				
st-KuNS	**К: Ааттарын һаҥаран көрүҥ диэ, кимнээгий?**				
ts-KuNS	– A:ttarɨn haŋaran körüŋ d͡ʲie, kimne:gij?				
tx-KuNS	A:ttarɨn	haŋaran	körüŋ	d͡ʲie,	kimne:gij?
mb-KuNS	a:t-tarɨ-n	haŋar-an	kör-ü-ŋ	d͡ʲie	kim-ne:g=ij
mp-KuNS	a:t-LArI-n	haŋar-An	kör-I-ŋ	d'e	kim-LA:K=Ij
ge-KuNS	name-3PL-ACC	speak-CVB.SEQ	try-EP-IMP.2PL	well	who-PROPR=Q
gg-KuNS	Name-3PL-ACC	sprechen-CVB.SEQ	versuchen-EP-IMP.2PL	doch	wer-PROPR=Q
gr-KuNS	имя-3PL-ACC	говорить-CVB.SEQ	попробовать-EP-IMP.2PL	вот	кто-PROPR=Q
mc-KuNS	n-n:poss-n:case	v-v:cvb	v-v:(ins)-v:mood.pn	ptcl	que-que>adj-ptcl
ps-KuNS	n	v	v	ptcl	adj
SeR-KuNS	0.3.h:Poss np:Th		0.2.h:A		
SyF-KuNS	np:O		0.2.h:S v:pred		
IST-KuNS	accs-inf		0.giv-inactive		
Top-KuNS					
Foc-KuNS	foc.int				
BOR-KuNS					
BOR-Phon-KuNS					
BOR-Morph-KuNS					
CS-KuNS					
fe-KuNS	– Try to name their names, who are they?				
fg-KuNS	– Versuchen Sie ihre Namen zu nennen, wer sind sie?				
fr-KuNS	— Назовите их имена, кто они?				
ltr-KuNS	К: Назовите их имена, кто они?				
nt-KuNS					

	43 [00:26.6]	44 [00:27.5]	45 [00:28.1]	46 [00:28.6]	47 [00:29.2]	48 [00:29.8]	49 [00:30.3]	50
ref-ElBK	ElBK_KuNS_2004_Storytellers_conv.ElBK.003 (001.007)							
st-ElBK	**E: Аата ким ити Ерёмин, тыа, каһуон олоҕкото баар, Александр..**							
ts-ElBK	– A:ta kim iti Jer'om'in, t͡ɨa, kah͡uon oloŋkoto ba:r, Al'eksandr…							
tx-ElBK	A:ta	kim	iti	Jer'om'in,	t͡ɨa,	kah͡uon	oloŋkoto	ba
mb-ElBK	a:t-a	kim	iti	Jer'om'in	t͡ɨa	kah͡uon	oloŋko-to	ba
mp-ElBK	a:t-tA	kim	iti	Jer'om'in	t͡ɨa	kah͡uon	oloŋko-tA	ba
ge-ElBK	name-3SG.[NOM]	who.[NOM]	that	Eryomin.[NOM]	Dolgan.[NOM]	some	tale-3SG.[NOM]	th
gg-ElBK	Name-3SG.[NOM]	wer.[NOM]	dieses	Jerjomin.[NOM]	Dolgane.[NOM]	einige	Märchen-3SG.[NOM]	es
gr-ElBK	имя-3SG.[NOM]	кто.[NOM]	тот	Еремин.[NOM]	Долган.[NOM]	несколько	сказка-3SG.[NOM]	ес
mc-ElBK	n-n:(poss)-n:case	que-pro:case	dempro	propr-n:case	n-n:case	quant	n-n:(poss)-n:case	pt
ps-ElBK	n	que	dempro	propr	n	quant	n	pt
SeR-ElBK	0.3.h:Poss np:Th						np:Th	
SyF-ElBK	np:S			n:pred			np:S	pt
IST-ElBK	giv-active			new			accs-inf	
Top-ElBK	top.int.concr							
Foc-ElBK				foc.nar		foc.wid		
BOR-ElBK				RUS:cult				
BOR-Phon-ElBK								
BOR-Morph-ElBK				dir:bare				
CS-ElBK								
fe-ElBK	– His name is Eryomin, a Dolgan, there are some tales of his, Aleksandr…							
fg-ElBK	– Er heißt Jerjomin, ein Dolgane, da sind einige Märchen von ihm, Aleksandr…							
fr-ElBK	— [Одного] зовут Ерёмин, долганин, несколько его сказок есть, Александр…							
ltr-ElBK	E: Зовут его Еремин, долганин, несколько его сказок есть, Александр…							
nt-ElBK								

Appendix B. Summary of notation conventions

Tiers	Notation	Meaning	See §
ts, tx	(word)	uncertain transcription	4.1.1
ts, tx	(wo-)	rejected fragment	4.2.2
ts, tx	(word=)	rejected word	4.2.2
ts, tx	((…))	unintelligible fragment	4.1.2
ts, tx	((wo…))	partially unintelligible fragment	4.1.2
ts, tx	((XYZ))	non-speech event	4.2.1
ts, tx	((XYZ:))	inline speaker change marker	2.4
ts, tx	((PAUSE))	long pause	4.2.1
ts, tx	…	unfinished sentence	2.2.1
ts, tx	. ? !	complete sentence	2.2.1
ts, tx	– —	turn-taking (sentence-initial)	2.4
ts, tx	-	orthographic morpheme boundary (compounds; clitics)	2.2.3
mb, mp	-	morpheme boundary	2.2.3
mb, mp	.[GLOSS]	zero morph	2.2.3
ge, gg, gr	%%	morpheme with unknown meaning	4.1.1
ge, gg, gr	%gloss	morpheme meaning uncertain	4.1.1
mc	%%	morpheme of unknown category	4.1.1
fe, fg, fr	(word?)	word with uncertain translation	4.1.1
fe, fg, fr	(word1/word2?)	two alternative translations	4.1.1
fe, fg, fr	word1 (/word)	word with an alternative translation	4.1.1
fe, fg, fr	word1 (/word?)	word with an alternative translation	4.1.1
fe, fg, fr	(…)	unintelligible fragment	4.1.2
fe, fg, fr	(?)	word with unknown translation	4.1.2
fe, fg, fr	[?]	sentence meaning unknown/uncertain (sentence-final)	4.1.3
fe, fg, fr	word1 [=word2]	literal translation [=intended/idiomatic translation]	4.3.4
fe, fg, fr	word1 [=word2?]	literal translation [=intended/idiomatic translation]	4.3.4
fe, fg, fr	[XYZ:]	inline speaker change marker (sentence-initial)	2.4
nt, nto	[XYZ:]	comment author	3.1