Uralic multimedia corpora: ISO/TEI corpus data in the project INEL

Timofey Arkhangelskiy
Universität Hamburg / Alexander von Humboldt Foundation
timarkh@gmail.com

Anne Ferger
Universität Hamburg
anne.ferger@uni-hamburg.de

Hanna Hedeland
Universität Hamburg
hanna.hedeland@uni-hamburg.de

Abstract

In this paper, we describe a data processing pipeline used for annotated spoken corpora of Uralic languages created in the INEL (Indigenous Northern Eurasian Languages) project. With this processing pipeline we convert the data into a lossless standard format (ISO/TEI) for long-term preservation while simultaneously enabling a powerful search in this version of the data. For each corpus, the input we are working with is a set of files in EXMARaLDA XML format, which contain transcriptions, multimedia alignment, morpheme segmentation and other kinds of annotation. The first step of processing is the conversion of the data into a certain subset of TEI following the ISO standard 'Transcription of spoken language' with the help of an XSL transformation. The primary purpose of this step is to obtain a representation of our data in a standard format, which will ensure its long-term accessibility. The second step is the conversion of the ISO/TEI files to a JSON format used by the "Tsakorpus" search platform. This step allows us to make the corpora available through a web-based search interface. As an addition, the existence of such a converter allows other spoken corpora with ISO/TEI annotation to be made accessible online in the future.

Tiivistelmä

Tässä paperissa kuvataan aineistonnprosessointimenetelmä joka on käytössä uralilaisten puhuttujen korpusten luonnissa kieltedokumentointiprojekti INELissä. Prosessointimenetelmää käytetään konvertoimaan dataa häviöttömään ISO/TEI- standardiformaattiin pitkän aikavälin säilytystä varten sekä samanaikaisesti tehokkaisiin hakutoimintoihin tälle akineistoversiolle. Jokaisen korpuksen lähtöaineistona on joukko tiedostoja EXMARaLDAn XML-formaatissa, joka sisältää transkriptejä,multimediaa kohdennuksineen, morfeemijäsennyksiä ja muita annotaatiota. Ensimmäinen käsittelyaskel on aineiston konvertointi TEI:n osajouk-

Proceedings of the fifth Workshop on Computational Linguistics for Uralic Languages, pages 115–124,
Tartu, Estonia, January 7 - January 8, 2019.

koon, joka muodostaa ISO-standardin puhutun kielen transkripteille, XSL- transformaatioita käyttäen. Tämän askelen ensisijainen tarkoitus on saada aineisto sellaiseen standardimuotoon joka kelpaa pitkäaikaissäilytykseen. Seuraava oaskel on ISO/TEI-tiedostojen konversio JSON-formaattiin, jota "Tsakorpus"-hakualusta käyttää. Tämän avulla saadaan korpus käytettäväksi internethakuliittymälle. Lisäksi, konversio mahdollistaa muiden ISO/TEI-yhteensopivien korpusten annotaatioiden tuomisen saataville tulevaisuudessa.

1 Introduction

The primary target of our processing pipeline are the corpora that are or will be developed within the framework of the INEL project (Indigenous Northern Eurasian Languages)[1] (Arkhipov and Däbritz, 2018). The main goal of the project is to develop annotated spoken corpora for a number of minority languages spoken in Northern Eurasia, most of them Uralic[2]. At the moment, corpora of Selkup (Uralic > Samoyedic), Kamas (Uralic > Samoyedic; extinct) and Dolgan (Turkic) are under development.

The long-term project INEL, scheduled to run until 2033, bases its technical development on the infrastructure, tools and workflows for curation and publication of digital resources available at the Hamburg Centre for Language Corpora[3], a research data centre within the CLARIN[4] infrastructure with a main thematic focus on spoken and multilingual data. The available technical solutions were however not developed for the specific data types created within the INEL project, in particular glossed and richly annotated transcripts. While the HZSK Repository used for corpus publication allows for transcript visualization and download, until now there is no advanced web-based search functionality. The INEL project thus needs to extend not only the existing workflows, but also the distribution channels to provide their corpora to a research community requiring easily accessible and highly complex search mechanisms.

Creating corpora from language data that are intended for long-term usage and accessibility holds various challenges for the used data formats. For each of the different phases the corpora undergo during their creation, different tools and therefore different data formats are needed. Because of the long-term character and the emphasis on accessibility of the INEL project, standard compliance and openness of the formats as well as making implicit information explicit also need to be taken into account. We will describe how we dealt with these challenges using a special data processing pipeline and using the ISO/TEI Standard Transcription of spoken language[5] (ISO/TC 37/SC 4, 2016) with only explicit information for short-term archiving, publishing, and as the base format for a searching interface.

Through the use of a standard format, the pipeline described in this paper can also be applied to similar (Uralic) corpora developed in other projects, e.g. the Nganasan Spoken Language Corpus (Wagner-Nagy et al., 2018).

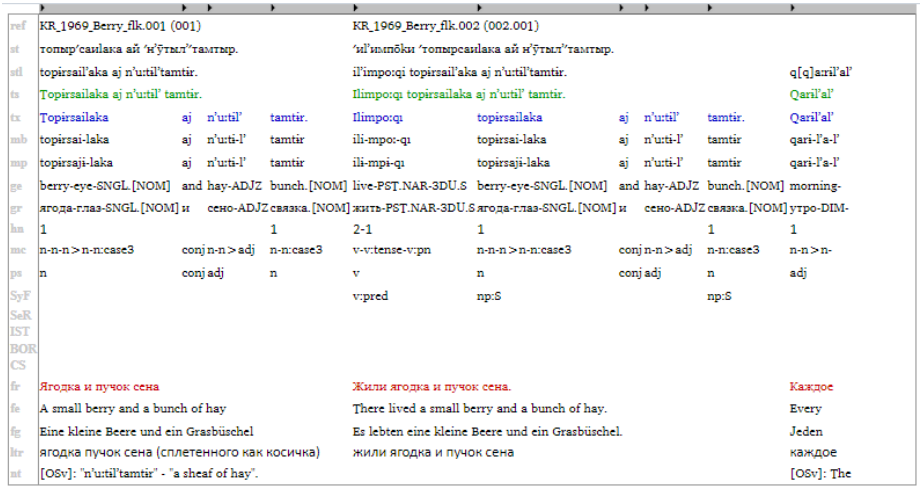
[1] https://inel.corpora.uni-hamburg.de/
[2] https://inel.corpora.uni-hamburg.de/?page_id=593
[3] https://corpora.uni-hamburg.de/
[4] https://www.clarin.eu/
[5] http://www.iso.org/iso/catalogue_detail.htm?csnumber=37338

2 Corpus Data

All corpora in question consist of audio/video files and/or pdf scans of (hand-)written data, accompanied by the transcription files. The annotation was carried out manually. It includes morpheme segmentation, glossing, part-of-speech tagging, syntactic functions, code switching, and other linguistically relevant information. When available, sound or video are aligned with the transcription. The files also contain parallel translations into Russian, English and German. While the transcription data is initially created with FLEx[6] and/or ELAN[7], the corpora within the INEL project are created with the EXMARaLDA[8] (Schmidt and Wörner, 2014) desktop software suite comprising a transcription and annotation editor (Partitur-Editor), a corpus and metadata manager (Coma) and a search and analysis tool (EXAKT). The transcriptions are stored in the EXMARaLDA Basic Transcription XML format (EXB) and have time-based references to the media files and multiple annotation tiers. Apart from the documentation on the creation of the corpus, detailed metadata regarding the raw data, the transcriptions and the corresponding speakers is stored in an additional file in the EXMARaLDA Coma XML format (COMA). All corpora in question have a similar design and similar annotation levels, which makes it possible to create a single set of converters capable of processing any of the existing corpora, as well as those that will be created in the course of the project. The EXMARaLDA software suite was chosen by the project because of the important advantages it offers for the corpus creation process. The metadata and corpus manager Coma and the desktop search and analysis tool EXAKT can both be used to manage and continuously assess the growing data set and also to search and analyze the corpus or any defined subcorpora. While the EXMARaLDA software might facilitate corpus creation, the time-based EXMARaLDA transcription data model is rather simple and not really suited for precise and explicit modelling of the INEL transcription data. The glossing comprises annotations which are clearly based on linguistic segments and such segment relations are not a part of the time-based EXMARaLDA transcription data model. The Basic Transcription format is not in any way tokenized and thus only allows for time-based annotations, aligned to start and end points shared with a transcription tier on which the annotation is based. In the INEL data, these start and end points coincide with token boundaries, though this is not explicitly modelled by the time-based approach. In Figure 1, you can see an example of a Selkup text annotated in the way described above.

The transcription files are generated either from (hand-)written text artifacts or from audio or video recordings. In case of transcriptions accompanied by audio or video, the EXMARaLDA transcriptions are aligned with these files by linking to time intervals in the media files. Some of those texts are dialogues. Since they are produced by multiple speakers, they contain several sets of tiers described above, one for each speaker. Finally, each corpus has very detailed metadata, which covers sociolinguistic background of the speakers and linguistic properties of the texts.

[6] `https://software.sil.org/fieldworks/`
[7] https://tla.mpi.nl/tools/tla-tools/elan/
[8] `https://exmaralda.org/de/`



ref	KR_1969_Berry_flk.001 (001)				KR_1969_Berry_flk.002 (002.001)					
nt	топыр'саилака ай 'н'ӱтыл''тамтыр.				'ил'илпӧки 'топырсаилака ай н'ӱтыл''тамтыр.					
utl	topirsail'aka aj n'u:til'tamtir.				il'impo:qi topirsail'aka aj n'u:til'tamtir.					q[q]aril'al'
ts	Topirsailaka aj n'u:til' tamtir.				Ilimpo:qı topirsailaka aj n'u:til' tamtir.					Qaril'al'
tx	Topirsailaka	aj	n'u:til'	tamtir.	Ilimpo:qı	topirsailaka	aj	n'u:til'	tamtir.	Qaril'al'
mb	topirsai-laka	aj	n'u:ti-l'	tamtir	ili-mpo:-qı	topirsai-laka	aj	n'u:ti-l'	tamtir	qari-l'a-l'
mp	topirsaji-laka	aj	n'u:ti-l'	tamtir	ili-mpi-qı	topirsaji-laka	aj	n'u:ti-l'	tamtir	qari-l'a-l'
ge	berry-eye-SNGL.[NOM]	and	hay-ADJZ	bunch.[NOM]	live-PST.NAR-3DU.S	berry-eye-SNGL.[NOM]	and	hay-ADJZ	bunch.[NOM]	morning-
gr	ягода-глаз-SNGL.[NOM]	и	сено-ADJZ	связка.[NOM]	жить-PST.NAR-3DU.S	ягода-глаз-SNGL.[NOM]	и	сено-ADJZ	связка.[NOM]	утро-DIM-
hn	1			1	2-1	1			1	1
mc	n-n-n>n-n:case3	conj	n-n>adj	n-n:case3	v-v:tense-v:pn	n-n-n>n-n:case3	conj	n-n>adj	n-n:case3	n-n>n-
ps	n	conj	adj	n	v	n	conj	adj	n	adj
SyF					v:pred	np:S			np:S	
SeR										
IST										
BOR										
CS										
fr	Ягодка и пучок сена				Жили ягодка и пучок сена.					Каждое
fe	A small berry and a bunch of hay				There lived a small berry and a bunch of hay.					Every
fg	Eine kleine Beere und ein Grasbüschel				Es lebten eine kleine Beere und ein Grasbüschel.					Jeden
ltr	ягодка пучок сена (сплетенного как косичка)				жили ягодка и пучок сена					каждое
nt	[OSv]: "n'u:til'tamtir" - "a sheaf of hay".									[OSv]: The

Figure 1: An example of a Selkup text annotated in EXMARaLDA

3 Conversion to ISO/TEI XML

While converters from the EXMARaLDA Basic Transcription format to the ISO/TEI standard are included in the EXMARaLDA software for several transcription conventions, these all assume that transcriptions were created with the time-based interpretation of EXMARaLDA and thus only create time-based annotations. With the ISO/TEI standard it is however possible to model segment based annotations, e.g. when further enriching the transcription data using tools such as parsers or taggers. For the INEL data, we decided to make the implicit information of the EXMARaLDA transcriptions, i.e. the segments and relations relevant to the annotation, explicit through corresponding modelling using ISO/TEI. The first objective was to convert the corpora into a loss-less standard format while turning implicit information into explicit information, which is especially important for long-term projects. Explicit data also means to make searching in various tools and converting into different data formats less error-prone. To achieve this we used the ISO standard "Transcription of Spoken Language"[9], which is based on the TEI guidelines[10]. To account for the specific requirements in the INEL project and similar structured projects (like Nganasan Spoken Language Corpus) we needed to use a defined subset of TEI that is segment-based and allows for segmentation into sentence, word and morpheme units while following the ISO standard. EXMARaLDA XML models transcription, description and annotation tiers as time-based information, linking these segments to the timeline of the linked media files. In the special case of the INEL corpora, there are also corpus files created in the EXMARaLDA format that don't reference any audio or video information because they are generated from (hand-)written text artifacts. The timeline in EXMARaLDA only needs to define events and not necessarily real time information, so in the text-based files references to those segments are used.

While this time-based format is needed for the transcription tier or "baseline", the annotations in INEL currently are exclusively segment-based, because they refer directly to the segments transcribed in the baseline tier and not the temporal events of

[9] http://www.iso.org/iso/catalogue_detail.htm?csnumber=37338
[10] http://www.tei-c.org/guidelines/p5/

the linked audio/video files. While still leaving the possibility for time-based annotations open in future work, we decided to convert the time-based annotations (that we knew to be segment-based) into segment-based annotations during the conversion into the standard ISO/TEI format, thus turning the implicit information we had into explicit information. The alignment of annotation segments to transcription tier segments can be deduced from a relation between the annotation and one transcription tier and their relation to events in the timeline. In the ISO/TEI conversion the annotations are linked directly to the segments of the baseline without additional time information. Depending on the scope of the annotations, annotation segments in the INEL TEI subset are linked to either sentence, word or morpheme units. Since these base units are linked to speakers (in EXMARaLDA as well as in the new ISO/TEI export), the annotation can be assigned to the respective speakers too.

The special subset of the ISO/TEI standard that we used contains automatically segmented `<seg>` elements (sentences) that consist of `<w>` elements (words) and punctuation elements `<pc>`. The annotations are structured into `<span>` elements. One special annotation tier additionally contains the morpheme segmentation, modelled by spans that have special ids. While there is an `<m>` element available in the ISO/TEI standard, we couldn't use it for our purposes, since `<m>` elements need to construct the words, while we need words additionally in a different tier than morphemes (which is needed for dealing with e.g. null morphemes of which we find many in our data). The start and end points of the spans can refer to the `<seg>` elements (sentences/utterances in the baseline), word elements or other spans (in our case: the morpheme spans). An sample fragment of a Selkup file in ISO/TEI can be seen in the Figure 2.

To account for the metadata, a metadata enriched search is planned in the future. To realize it, the metadata from the Coma XML file concerning the transcriptions and speakers will be exported in the ISO/TEI files during the conversion additionally, using the `<tei:teiHeader>`.

4 Conversion from ISO/TEI to Tsakorpus

Our second objective was to give the linguists access to the corpora through a user-friendly web-based interface. We use the "Tsakorpus" corpus platform for this. In this platform, linguistic data is indexed in a document-based Elasticsearch database. The main index contains sentences (or sentence-like sequences of tokens), each sentence being a single document. Tsakorpus accepts files in a certain JSON-based format as its input. Each file corresponds to one corpus text (i.e. one EXMARaLDA transcription file in our case) and contains a list of sentences. A sentence contains a list of tokens, which are also represented as JSON objects, and additional information such as sentence-level metadata, audio/video alignment and parallel alignment.

In order to index our corpora in Tsakorpus, we wrote a Python script that transforms ISO/TEI files to the JSON files required by the corpus platform. Since all information potentially relevant for the search is already explicit in the TEI files, the conversion basically means simply recombining the existing data without recovering the information stored implicitly (such as grammatical values expressed by null morphemes). The TEI files have a tier-based structure. This means that for each property (such as part of speech, morpheme segmentation etc.), its values are listed for every word within an XML node representing that property, and nodes representing different properties follow one another. In Tsakorpus JSON, properties of each token are all

stored together in the JSON object representing that token. An example featuring a fragment of a source file and the resulting JSON tokens can be seen in 3 and 4. Several tier names commonly used by linguistic annotation software (FLEX or Toolbox) are translated into names reserved for certain annotation types in Tsakorpus. E.g. the information from the `ps` tier, which represents part of speech, goes to the `gr.pos` tier in the JSON. All unrecognized tier names, such as `SyF` in the example (syntactic function) are left unchanged.

There are three processing steps that go beyond simple restructuring described above.

First, information about the alignment with the media file should be included in each sentence. In the ISO/TEI files, the alignment is indicated through time point labels such as `T3`, which come directly from the EXMARaLDA files. In the beginning of the ISO/TEI document, all these labels are listed with their time offset values in seconds. The name of the media file (one per transcription) associated with the recording is stored in the `sourceDesc` node of the ISO/TEI file.

The time point labels at sentence boundaries are replaced with the actual time offsets in the JSON. Additionally, the source media file is split into overlapping pieces of small length (60 seconds by default) using `ffmpeg`. Instead of being associated with the entire media file, each sentence in the JSON file is associated with one of these parts. The part for each sentence is chosen in such a way that the middle of the sentence segment is as close as possible to the middle of the media segment. The time offsets are changed accordingly. Such an approach allows the user to listen to the segment they found together with some context, while at the same time avoiding the need to download the entire media file, which could be quite large.

Second, there is a number of tiers with sentence-level alignment in the source files. These are alternative transcriptions and translations into Russian, English and German. To enable this sort of alignment in Tsakorpus, we are using a scheme intended for parallel corpora. The aligned segments are stored in the JSON file as separate sentences. Sentences originating in different tiers have different values of the `lang` parameter. The sentences that should be aligned with one another receive the same "parallel ID", a value stored in each of them.

Finally, the translations are automatically lemmatized and morphologically analyzed using the analyzers available for Russian, English and German. As of now, we have tested the analysis of the Russian tier with `mystem` (Segalovich, 2003). This may seem a significant departure from the principle of having all relevant information explicitly present in the ISO/TEI files. However, we treat this added annotation as an auxiliary information that is not part of the original annotated corpus and should not be stored in it. Its only purpose is facilitating the search in the data that already exists in the corpus. The queries that involve this annotation are not intended to be replicable. Therefore, this annotation is not checked manually and can be superseded by annotation produced by other morphological analyzers in the future.

After the JSON files are indexed, the corpus becomes available through a web interface. Single-word queries may contain constraints on values in any annotation tier or their combinations, possibly with regular expression or Boolean functions. Multiword queries can additionally include constraints on the distance between the words. Each search hit is aligned with a multimedia segment, which can be played by clicking on the sentence.

5 Conclusion

By developing the EXMARaLDA > ISO/TEI and ISO/TEI > Tsakorpus JSON converters we have achieved two goals. First, the corpora annotated within the framework of INEL and similar projects can now be exported to a format suitable for long-term preservation. The version of the ISO/TEI we are using is fit for that purpose because it is based on an ISO standard and because all potentially relevant information is made explicit in it. This means that the corpora in question could be reused in the future without recourse to the software currently employed in the project or to implicit knowledge of its participants. Second, this chain of converters makes it possible to release the corpora to the public through a user-friendly web interface. This way of publishing the corpora has an advantage over simply releasing the EXMARaLDA files in that it does not require the users to install and become acquainted with any special software.

The ISO/TEI > Tsakorpus JSON converter is open source, which means that any corpus stored in a similar ISO/TEI form could be easily published online. Projects that use ISO/TEI for storing annotated spoken corpora exist, e.g. in IRCOM infrastructure (Liégeois et al., 2015), but are not numerous. The ISO/TEI format is aimed at creating enhanced interoperability for spoken data through a standardized format. Apart from the proof of concept work done by integrating transcription data from various tool formats into an ISO/TEI corpus that can be searched in its entirety, support for various other scenarios, such as linguistic web services and web-based annotation tools are in development.

Importantly, the availability of our converter could encourage researchers working in language documentation projects to export their data to ISO/TEI, which would be beneficial for their long-term availability.

Acknowledgments

This publication has been produced in the context of the projects CLARIN-D, funded by the German Ministry for Education and Research (BMBF) under grant number 01UG1620G, and INEL, within 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.

References

Alexander Arkhipov and Chris Lasse Däbritz. 2018. Hamburg corpora for indigenous Northern Eurasian languages. *Tomsk Journal of Linguistics and Anthropology* (3):9–18. https://doi.org/10.23951/2307-6119-2018-3-9-18.

ISO/TC 37/SC 4. 2016. Language resource management – Transcription of spoken language. Standard ISO 2462:2016, International Organization for Standardization, Geneva, CH. http://www.iso.org/iso/catalogue$_detail.htm?csnumber = 37338$.

Loïc Liégeois, Carole Etienne, Christophe Parisse, Christophe Benzitoun, and Christian Chanard. 2015. Using the TEI as a pivot format for oral and multimodal lan-

guage corpora. Paper presented at Text Encoding Initiative Conference, Lyon, October 28–31, 2015.

Thomas Schmidt and Kai Wörner. 2014. EXMARaLDA. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, Oxford University Press, pages 402–419. http://ukcatalogue.oup.com/product/9780199571932.do.

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA-2003*. Las Vegas.

Beáta Wagner-Nagy, Sándor Szeverényi, and Valentin Gusev. 2018. User's Guide to Nganasan Spoken Language Corpus. *Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology* 1.

```
<seg subtype="declarative" xml:id="seg1">
    <w xml:id="w1">Qwärɣa</w>
    <anchor synch="T1" />
    <w xml:id="w2">t’üumbädi</w>
    <anchor synch="T2" />
    <w xml:id="w3">surɨm</w>
</seg>
...
<spanGrp type="st">
    <span from="seg1" to="seg1">ˊkwäpχa ˊ
    тˊўмбäди ˊсурым.</span>
...
</spanGrp>
<spanGrp type="mb">
     <span from="w1" to="w1">
            <span xml:id="m1">qwärɣa</span>
     </span>
     <span from="w2" to="w2">
            <span xml:id="m2">t’üu</span>
            <span xml:id="m3">mbädi</span>
      </span>
...
</spanGrp>
<spanGrp type="ge">
      <span from="w1" to="w1">
            <span from="m1" to="m1">bear</span>
            <span>NOM</span>
      </span>
      <span from="w2" to="w2">
            <span from="m2" to="m2">get.angry</span>
            <span from="m3" to="m3">PTCP.PST</span>
       </span>
...
</spanGrp>
...
<spanGrp type="ps">
       <span from="w1" to="w1">n</span>
        ...
</spanGrp>
```

Figure 2: Segment-based morpheme-segmented Subset of ISO/TEI

```
<seg subtype="declarative" type="utterance" xml:id="seg1">
    <w xml:id="w4">It'e</w>
    <w xml:id="w5">pal'd'ukus</w>
</seg>
...
<spanGrp type="ps">
    <span from="w4" to="w4">nprop</span>
    <span from="w5" to="w5">n</span>
</spanGrp>
<spanGrp type="SyF">
    <span from="w4" to="w4">np.h:S</span>
    <span from="w5" to="w5">v:pred</span>
</spanGrp>
```

Figure 3: Tier-based data representation in ISO/TEI

```
{
 "wf": "It'e",
 "wtype": "word",
 "ana": [
  {
   "gr.pos": "nprop",
   "SyF": "np.h:S"
  }
 ],
 "off_start": 0,
 "off_end": 4
},
{
 "wf": "pal'd'ukus",
 "wtype": "word",
 "ana": [
  {
   "gr.pos": "n",
   "SyF": "v:pred"
  }
 ],
 "off_start": 5,
 "off_end": 15
}
```

Figure 4: Token-based data representation in Tsakorpus JSON