Northern European Journal of Language Technology, 2016, Vol. 4, Article 3, pp 29–47
DOI 10.3384/nejlt.2000-1533.1643

Utilizing Language Technology in the Documentation of Endangered Uralic Languages

Ciprian Gerstenberger[*]
UiT – The Arctic University of Norway
Giellatekno – Saami Language Technology
ciprian.gerstenberger@uit.no

Niko Partanen
University of Hamburg
Department of Uralic Studies
niko.partanen@uni-hamburg.de

Michael Rießler
University of Freiburg
Department of Scandinavian Studies
michael.riessler@skandinavistik.uni-freiburg.de

Joshua Wilbur
University of Freiburg
Department of Scandinavian Studies
joshua.wilbur@skandinavistik.uni-freiburg.de

5 January 2017

Abstract

The paper describes work-in-progress by the Pite Saami, Kola Saami and Izhva Komi language documentation projects, all of which record new spoken language data, digitize available recordings and annotate these multimedia data in order to provide comprehensive language corpora as databases for future research *on* and *for* endangered – and under-described – Uralic speech communities. Applying *language technology* in *language documentation* helps us to create more systematically annotated corpora, rather than eclectic data collections. Specifically, we describe a script providing interactivity between different morphosyntactic analysis modules implemented as Finite State Transducers and ELAN, a Graphical User Interface tool for annotating and presenting multimodal corpora. Ultimately, the spoken corpora created in our projects will be useful for scientifically significant quantitative investigations on these languages in the future.

[*] The order of the authors' names is alphabetical.

1 Introduction

Language documentation (aka documentary linguistics) is an emerging sub-field of applied linguistics. Research in language documentation aims at the provision of long lasting, comprehensive, multi-faceted and multi-purpose records of linguistic practices characteristic of a given speech community (cf. Himmelmann 2006; Woodbury 2011; Austin 2014). Although it evolved out of traditional fieldwork methodology used primarily by descriptive linguists and language anthropologists, language documentation is no longer merely a method, as it has its own primary aims and methodologies. One of the most important purposes of language documentation is making data available for further research on and for endangered languages, for both further theoretical and applied research, as well as for direct use by the relevant language communities.

Ideally, the data pool provided by the language documenter includes a comprehensive, deeply annotated and easily accessible corpus of primary language recordings, representing a wide variety of texts in terms of chronology (e.g. age of recorded speakers), geography (e.g. dialects), and other sociolinguistic variables (e.g. gender and educational background of speakers, registers, genres, etc.). In addition to annotations, cataloging metadata are crucial for the intellectual accessibility of the documented data and concern both the *content* of the recorded speech sample (typically represented as phonological transcriptions, morphosyntactic glossing and tagging, and translations) as well as the *context* (such as actors, places, speech events, but even meta-documentation about the project itself, cf. Austin 2013).

Along with methodologies and best practices related to fieldwork and archiving (including questions of research ethics, protection of copyrights, resource discoverability, data standards and long term data safety), the usefulness of the actual product of language documentation for linguistic research hinges on the quality and quantity of *annotations* as the basis for further analyses and data derivations. The use of language documentations for corpus-based investigations on endangered and less-known languages and the role of computational linguistics for the field has frequently been a driving topic in recent years.

In fact, with respect to the data types involved, endangered language documentation generally seems similar to language corpus building, at least in principle. Both provide primary data for secondary (synchronic or diachronic) data derivations and analyses (for data types in language documentation, see Himmelmann 2012). The main difference is that traditional corpus and computational linguistics deals predominantly with larger non-endangered languages, for which huge amounts of mainly written corpus data are available. The documentation of endangered languages, on the other hand, typically results in rather small corpora of exclusively spoken genres. Furthermore, corpus annotations in language documentation projects are often created manually. As a result, significant quantitative investigations normally do not make use of spoken corpora from endangered language documentation projects.

Regarding collaborative tools and user interfaces for transcribing, archiving and browsing multimedia recordings, language documentation has made huge technological progress. However, paradoxically, the field has only rarely considered applying automated methods to annotate data more efficiently – both with respect to quality and to quantity – in order to create a solid foundation for new and better corpus-based linguistic research on smaller languages. Relatively small endangered languages are becoming more and more the focus of computational linguistic research, and relevant language-technological methods and tools are completely functional even for relatively small languages such as Northern Saami, Komi-Zyrian or Komi-Permiak today (e.g. Trosterud 2006a for Saami languages and Snoek et al. 2014 for Plains Cree; see also Poibeau and Fagard 2016 for a different approach). Nevertheless, these methods and tools are still being applied exclusively in corpus-building of *written* language varieties.

Current language technology projects on endangered languages (including the Giellatekno group,[1] which we are both affiliated with and enjoy a significant collaboration with) seem to have simply copied their approach from already established research on larger, non-endangered languages, including the focus on written language. The resulting corpora are impressively large for these minority languages and include higher-level morphosyntactic annotations. However, they represent a limited range of text genres, and typically include mainly formal styles, while also consisting to a large extent of translations from the relevant majority languages.[2]

Note also that, as the current written standards of small endangered languages, like Northern Saami, are to a large part evolving as the result of institutional language planning, the bulk of texts in the Northern Saami corpus consists of translations from the majority languages. Even original Saami texts (e.g. from official webpages and the few newspapers) are most typically produced only by a few writers.[3]

The restriction on written language is even more crucial in the case of smaller languages for which language technology is currently under active development, such as for Skolt Saami.[4] Although active language planning for Skolt Saami was already initiated several decades ago and the amount (and quality) of written texts is ever growing, the language is still most typically only used in speech. As a result, there is an urgent need to enrich the existing corpora for languages such as Northern Saami and Skolt Saami with new data from spoken genres.

For exceptionally small Saami languages such as Pite Saami, the texts available for corpus creation are almost exclusively in a non-written modality, and an efficient and consistent method for incorporating spoken texts is vital for corpus creation. In fact, spoken language documentations for such languages often exist and several projects continue collecting new recordings and annotating legacy speech samples. However, as much as endangered language documentation and language technology seem to overlap in their respective general agendas towards applied linguistic research, both fields have scarcely

[1] `http://giellatekno.uit.no`

[2] The metadata provided with the Northern Saami written corpus at Giellatekno (Giellatekno and Divvun 2016) suggests that the portion of non-translated texts is in fact rather high, which refutes our previous claim in Blokland, Gerstenberger, et al. (2015).

[3] In the case of existing corpora for written Northern Saami there seems to be one more problematic bias in that so far work with corpus building has been carried out almost exclusively by the Norwegian based Giellatekno initiative, focusing on texts produced in Norway and hence representing only the varieties of written Northern Saami used in Norway. Texts produced in Finland and Sweden are only marginally included in Giellatekno's Northern Saami corpora.

[4] For Skolt Saami, see especially the project *Koltansaamen elvytys kieliteknologia-avusteisen kielenoppimisohjelmien avulla sekä mallin ja ohjeiden laatiminen menetelmän siirtämiseksi toisiin uhanalaisiin kieliin* carried out by Jack Rueter and collaborators at the University of Helsinki, see `http://www.koneensaatio.fi/rohkeat-avaukset/tuetut/tuetut2014/kielten-elvyty/`, and in collaboration with Giellatekno and the Skolt Saami language community in Finland.

met so far.

Compared to the computational linguistics projects described above, researchers working in the framework of endangered language documentation, i.e. fieldwork-based documentation, preservation, and description of endangered languages, often collect and annotate natural texts from a variety of spoken genres and including both formal and informal styles. Commonly, the resulting spoken language corpora have phonemic transcriptions as well as several morphosyntactic annotation layers produced either manually or semi-manually with the help of software like Field Linguist's Toolbox (or Toolbox, for short),[5] FieldWorks Language Explorer (or FLEx, for short)[6] or similar tools. Common morphosyntactic annotations include glossed text with morpheme-by-morpheme interlinearization. Whereas these annotations are qualitatively rich, including the time alignment of annotation layers to the original audio/video recordings, the resulting corpora are relatively small and rarely reach 150,000 word tokens. One example of a comparably large corpora created in this approach, and supposedly even exceeding the number of 150,000 tokens, is the corpus of Forest and Tundra Enets (Comrie et al. 2005–2017).[7] The Nganasan corpus described in Beáta and Szeverényi 2015 totals approximately 100,000 tokens.[8] Typically, such spoken corpora are smaller, as is the case with the annotated corpora of Tundra Nenets and Northern Khanty.[9] The main reason for the limited size of such annotated language documentation corpora is that (semi-)manual glossing is an extremely time consuming task.

Another problem we identify especially in the documentation of small Uralic languages is that projects sometimes ignore the existence of orthographies and prefer phonemic or even detailed phonetic transcription.

Note that most Uralic languages (or at least their main variants) have established written standards as the result of institutionalized and/or community-driven language planning and revitalization efforts. For some of these languages, e.g. Northern Khanty, Komi-Zyrian, Northern Selkup, Tundra Nenets or Udmurt, a significant amount of printed texts can be found in books and newspapers[10] and several of these languages are also used digitally on the Internet today[11] which makes it possible to combine spoken and written data in one and the same corpus.

Last but not least, there are at least small dictionaries available for all of these languages, several of which have already been digitized. The use of such lexical materials for the purpose of automatic corpus annotation has even been reported to be quite successful (Arkhangelskiy and Medvedeva 2016).

Particularly when basic phonological and morphological descriptions are already avail-

[5] http://www-01.sil.org/computing/toolbox

[6] http://fieldworks.sil.org/flex

[7] Olesya Khanina, p.c.; an indication of the actual size, in terms of texts, sentences, tokens or the like, is otherwise not given.

[8] Valentin Gusev, p.c.; note that this corpus overlaps to a great extent with the one presented on the site http://www.iling-ran.ru/gusev/Nganasan.

[9] http://larkpie.net/siberianlanguages/recordings/tundra-nenets and http://larkpie.net/siberianlanguages/northern-khanty; an indication of the actual size, in terms of texts, sentences, tokens or the like, is not given.

[10] For printed sources from the Soviet Union and earlier, the Fenno-Ugrica Collection is especially relevant: http://fennougrica.kansalliskirjasto.fi; contemporary printed sources are also systematically digitized, e.g. for both Komi languages: http://komikyv.ru.

[11] See, for instance, The Finno-Ugric Languages and The Internet Poject (Jauhiainen et al. 2015).

able and can serve as a resource for accessing phonological and morphological structures (which is arguably true for the majority of Uralic languages), we question the special value given to time-consuming phonemic transcriptions and (semi-)manual morpheme-by-morpheme interlinearization. Instead, we propose a step-by-step approach to reach higher-level annotations by using and improving truly computational methods, while systematically integrating all available textual, lexicographic, and grammatical resources into the language documentation endeavor (see also Blokland, Gerstenberger, et al. 2015).

2 Language Documentation Meets Language Technology

Our projects are concerned with the building of multimodal language corpora, including at least spoken and written (i.e. transcribed spoken) data and applying innovative methodology at the interface between endangered language documentation and endangered language technology. We understand language technology as the functional application of computational linguistics as it is aimed at analyzing and generating natural language in various ways and for a variety of purposes. Machine-based translation or automatic language analyzers and morphosyntactic taggers are but two examples of such practical applications. We believe that all combined efforts between language technology and language documentation can clearly be directly profitable both for corpus-based theoretical investigations and for language planning and revitalization of endangered languages. Whereas the language documenters provide the speech corpora and linguistic analyses necessary for the computational modeling of the languages in question, language technologists apply formal-descriptive linguistic and corpus linguistic methods for the programming of machine-readable grammatical and lexical descriptions of the relevant languages in order to create computer tools for language users. Spoken language documentations can thus increase the size of the data pool utilized in the research carried out by computational linguists and language technologists. Language technology, on the other hand, can create tools which analyze spoken language corpora in a much more effective way, and thus allow one to create better linguistic annotations for and descriptions of the endangered languages in question. This allows for directing more resources towards transcription and for working with larger data sets because slow manual glossing no longer necessarily forms a bottleneck in a project's data management workflow.

Here, we describe our current work on recording and annotating spoken language data and discuss the combination of methods from both language documentation and language technology used by our projects. At present, we are working with the following languages: Pite Saami, Skolt Saami, Kildin Saami[12] as well as the Izhva and Udora varieties of Komi-Zyrian[13] (cf. also Wilbur 2008–2017; Rießler 2005–2017; Blokland, Fedina, et al. 2009–2017; Partanen et al. 2013). Using data examples from our current projects, we will show how *language documentation* can profit significantly from the application of automated corpus data annotation, specifically the application of rule-based morphosyntactic tagging for our spoken corpora. *Language technology*, on the other hand, can profit from the use

[12] http://saami.uni-freiburg.de

[13] http://fu-lab.ru

of more extensive and more diverse data, which our projects provide and which include predominantly natural spoken language data.

We suggest the following two main principles, which we have begun implementing consistently in our own documentation projects:

1. use an orthography-based transcription system, and

2. apply computer-based methods as much as possible in creating higher-level annotations of the compiled corpus data.

In addition to designing annotation schemata of appropriate granularity for corpus building, two essential aspects of language documentation remain important in our approach: the *archiving* of primary data linked to all data derivations, as well as proper *contextualization* by means of DEEP METADATA concerning the non-linguistic context of a given speech sample. By 'DEEP METADATA' we mean metadata concerning a variety of levels of sociolinguistic, pragmatic and other descriptions in addition to basic cataloging facts (such as time and place of a recording). Computational and corpus linguistic approaches to applied research on endangered languages (including Giellatekno) have scarcely considered the latter aspects, even though these are crucial for language documentation aimed at long-lasting, comprehensive, multi-faceted and multi-purpose records of linguistic practices. Our approach has also proven challenging for current practices within language documentation. The majority of metadata is related to complete sessions, which is also the level at which all metadata models such as IMDI and CMDI operate (for a brief discussion on metadata formats, see Section 4.1). However, it is also possible to associate partial events with their own metadata. This is useful for example when the recording was done outside and the recording location and context actually change throughout the session. Similarly, one recording may be split into several stories, narratives and songs, which each inherit the majority of the session metadata, but may each also have more specific descriptive information. No conventions or best practices for working with metadata on this level currently exist.

Below, we present our work-in-progress concerning the application of rule-based morphological tagging in automatically creating corpus annotations. The next step is then to take the annotation work further to syntactic disambiguation and on to parsing. In this, our aim is to challenge, further develop and extend current approaches at the interface between computational, descriptive and documentary linguistics for endangered languages.

3 From Fieldwork to Corpus Data

The following section describes the entire workflow of corpus data collection, processing and archiving. However, our description focuses on the higher-lever linguistic annotation of the data (see Section 4.1). Section 3.1 provides only a general overview of the origin and preliminary processing of our fieldwork data as well as its long-term archiving. Since this paper concerns the computational aspects of linguistic corpus annotation specifically, we refrain from describing specific workflows for anthropological linguistic fieldwork and language archiving or standards for multimedia language data. Our general conventions are similar to other endangered language documentation projects; Gippert et al. 2006

Table 1: Overview on the amount of data in our projects at present; the category `Tokens` refers to the number of transcribed tokens in both audio/video recordings and digitized transcribed spoken texts lacking a recording; note that these numbers are only very rough estimates; typically our data also include translations into at least one majority language

Language	**ISO code**	**Recorded speakers**	**Time span of texts**	**Tokens**
Pite Saami	sje	17	1893–2016	27,000
Skolt Saami (Notozero dialect)	sms	12	1876–2016	21,000
Akkala Saami (Babino dialect)	sia	9	1971–1987	3,000
Kildin Saami	sjd	~70	1876–2014	110,000
Ter Saami	sjt	~20	1856–2006	8,000
Komi-Zyrian (Izhva dialect)	kpv	~150	1844–2016	200,000
Komi -Zyrian (Udora dialect)	kpv	~50	1902–2013	40,000

can serve as a general reference, while Wilbur 2014 and Wilbur 2011 provide specific examples from the documentation of Pite Saami.

3.1 Collection and Preliminary Processing of Data

The bulk of speech data we include in our corpora originates from our own fieldwork recordings. We use high quality audio recording devices – if possible in combination with a video recorder – and typically more than one external microphone. This results in several audio and video tracks which have to be combined into a single session file before further processing the resulting data and their inclusion into our corpora.

The segmentation of audio/video recordings from our own fieldwork and their transcription and translation is done in ELAN (see Section 4.1). Depending on the project, these tasks are often shared by several collaborators (in this, native speaker collaborators are usually responsible for transcribing and/or translating the respective languages), but sometimes all necessary working steps are finished by a single individual. Since surface text structuring in spoken texts is prosodic rather than syntactic, we use utterances as the basic units of our text segmentation in ELAN, rather than sentences. Nevertheless, we represent our spoken recordings using standardized orthography (with adaptations for dialectal and other sub-standard forms as needed), rather than phonemic transcription – unlike many other endangered language documentation projects. We also use orthographic punctuation marks for clause or phrase boundaries, similar to written language.

As described above, one of our main principles is using an orthography-based transcription system. This not only allows for quicker and more efficient transcription of field recordings by our native speaker collaborators, who are more used to writing orthography than a writing system based on a transliteration or a phonetic/phonemic transcription. Using orthography also makes it possible to easily integrate all available (digitized) printed texts into the corpus (see Section 3.2 for more on legacy data). In addition, any available (digitized) lexical resources can be integrated into the annotation tools under creation as well, rather than building new dictionaries from scratch via interlinearization (which is the typical approach by projects using FLEx or similar tools). Note, however, that not all digitized texts or dictionaries are written using the standard orthography. In particular, the samples of spoken language collected by linguists and printed as transcriptions (and translations) in books typically use (Latin or Cyrillic) non-orthographic writing systems, such as the Finno-Ugric transcription system. The same is true for descriptive dictionaries targeting a scientific audience. While we keep the original transcriptions after digitization whenever feasible, normalizing these into standard orthography is the most effective way to integrate all different spoken texts into a single writing system and then to combine these texts with available written corpora, which of course use standard orthography.

3.2 Legacy Data

In addition to our own field work data, we include available legacy data in our corpora whenever possible. By 'LEGACY DATA' we mean for instance fieldwork data collected by other projects (annotated or not) and stored in various language archives, as well as spoken texts which were transcribed, translated and published in books and are available with or without any original recordings files. Further processing of legacy data basically follows the same processing as with our fieldwork data, and thus includes segmentation into utterances in ELAN, followed by orthographic transcription (potentially including conversion or correction of an already existing transcription) and translation. Even spoken texts samples published in books are further processed in ELAN after digitization (if they are not available in digital form already).

3.3 Archiving and Publication

Our projects' language documentations are archived at The Language Archive (TLA)[14] for Kola Saami and Komi data and The Endangered Languages Archive (ELAR)[15] for Pite Saami data. In addition, the Pita Saami data is mirrored in TLA. Both archives provide online user-interfaces to browse and access the data as well as define access rights.

In the archives, metadata stored in CMDI[16] format is linked to each ELAN annotation file in order to keep track of situational or contextual factors that are related to the data in one way or another. For instance, in order to preserve more pieces of information about the sessions, details about different speakers, the recording setting, or the recording devices used, as well as about work with specific projects or individuals can be included in the metadata. It is also desirable to store metadata separately from basic annotations because this makes it very easy to control access to more sensitive pieces of information that may be stored in the metadata. In our model, individual session names and anonymizable actor IDs can be used to associate any (sub-)set of metadata with any transcription.

Currently, we are already able to carry out searches on a subset of the corpus by using metadata constraints, for instance, on participants' ages or regional affiliations.

[14] `https://tla.mpi.nl`

[15] `http://www.elar-archive.org`

[16] For CMDI (`Component MetaData Infrastructure`) format, see `http://www.mpi.nl/cmdi`.

This provides a solid fundament for quantitative and more fine-grained sociolinguistically oriented research. In principle, the transcription files also contain small traces of metadata since the filenames are standardized to include the language ISO-code and the recording date.

In addition to the archives mentioned above, our data is available (or will be made available in the near future) though other user portals, such as an interface called Video Corpora[17] maintained by our project partner FU-Lab in Syktyvkar which mainly targets the native speaker community, although the available search functions and the representation of the data in three different meta-languages make this multimedia text collection useful for linguists and other users from outside the speaker communities as well.

A possible corpus interface specifically for linguists interested in our data is Korp,[18] which is already being used by Giellatekno for Saamic and other languages. However, Korp is not currently able to link corpus data to audio/video media.

4 Corpus Data Processing

In this section, we give a detailed account of the structure of an ELAN file as used in our projects. Although we present the ELAN Graphical User Interface (GUI) as a tool for manual annotation, the emphasis lies on the description of the pipeline that allows us to extract word forms from the ELAN files, send them to an finite state transducer (FST) for morphosyntactic analysis, and add these analyses in the correct place in the ELAN XML structure.

4.1 ELAN as a Tool for Annotating Multimedia Data

Our language corpora represent spoken and written text modi of formal and informal registers and a variety of genres. Our transcribed spoken text data (using standard orthography) as well as any written text data are stored in XML format and structured to be utilized by the multimedia language annotation program ELAN.[19] This software allows audio and video recordings to be time aligned with detailed, hierarchically organized tiers for transcriptions, translations and further annotations (cf. Figure 3). Furthermore, using ELAN as a corpus search tool, basic frequency statistics can be calculated, concordances created, and data for statistical analysis exported (e.g. using R[20] or similar tools).

ELAN annotation tiers used in our projects are organized hierarchically based on the minimal template shown in Figure 1 for each speaking participant in a recording. Since each speaker has his/her own time-aligned tier node `ref`, including dependent tiers, annotating simultaneous speech by multiple speakers (a common feature of spoken language) is not problematic. Adding new tiers via the ELAN GUI is very easy; this way one can, for instance, collect different orthographic standards or transcription systems used in different publications for the same text. For older materials it is not uncommon

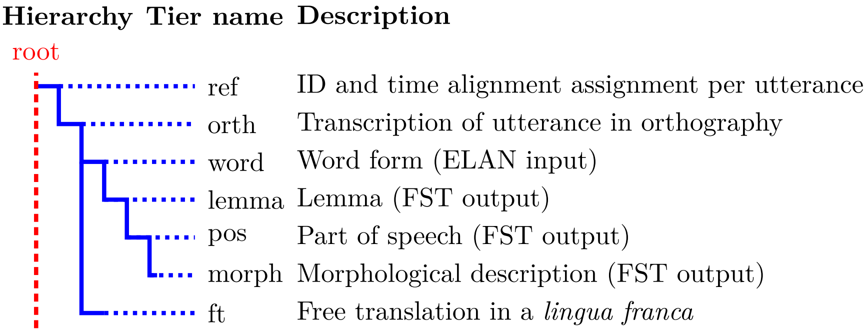
[17] http://videocorpora.ru

[18] http://gtweb.uit.no/korp

[19] ELAN (EUDICO Linguistic Annotator) is free software developed by the Technical Group of the Max Planck Institute for Psycholinguistics, cf. https://tla.mpi.nl/tools/tla-tools/elan.

[20] http://www.r-project.org; see Nagy and Meyerhoff 2015 for the use of ELAN for coding and extracting corpus-linguistic data for statistical analyzes.

that they have been already published different times, but usually the work undertaken in our projects reflects the first time the data are associated with the original audio.



Hierarchy	Tier name	Description
	ref	ID and time alignment assignment per utterance
	orth	Transcription of utterance in orthography
	word	Word form (ELAN input)
	lemma	Lemma (FST output)
	pos	Part of speech (FST output)
	morph	Morphological description (FST output)
	ft	Free translation in a *lingua franca*

Figure 1: The basic ELAN tier hierarchy used in the documentation corpora described here

While ELAN is intended mainly as a time-aligned interface between written annotations (transcriptions, translations and others) and the original audio/video medium, it is also possible to use ELAN for texts only, i.e. for written texts without audio/video. In this way, legacy texts of speech recordings without multimedia[21] as well as pure text data are also included in the corpora we create.

4.2 ELAN as a Corpus Tool

It is possible to execute complex searches on multiple ELAN files – on the entire corpus or only specific parts of it. Search constraints on the name or type of tier, contextual information, etc., while using regular expressions, can be specified in various combinations. The search facilities use an ELAN file's tier-based structure very conveniently, allowing one to search for co-occurrences of items on different, but connected tiers. Yet, with this data alone it is not possible to filter results by adding more contextual constraints. ELAN is currently able to display session related metadata from CMDI files, but it is not possible to use this metadata in searches conducted within ELAN. Such filtering is thus only possible using the associated metadata in some other environment. As an example, if one wants to filter out the tokens produced by non-native speakers, or include only those speakers who are born in a specific region, the search or later filtering of search results has to performed outside ELAN.

Search results can be shown in a *key-words in context* (KWIC) format, i.e. in a concordance in which up to eight words on either side of the search term are visible. A complex multiple-layer search limiting the context of neighboring or hierarchically related tiers can also be performed. The basic search functionalities can be compared to other corpus interfaces, but ultimately the annotations (the data themselves) are the most important factor in determining what kinds of search are feasible. As for exporting, all search results can be saved in plain text in comma-separated values (CSV) format. Finally,

[21] Some examples are Lagercrantz' *Lappische Volksdichtung* published between 1957–1966, which includes transcribed recordings for both Pite and Skolt Saami, as well as numerous similar text samples of other endangered Uralic languages for which fieldworkers transcribed speakers phonetically without using a recording device.

the ELAN format is an XML format (with the file extension `.eaf`), and as such it is archive-friendly but also somewhat human-readable. The format will likely be supported well into the near future as XML is a common and open-source format. In many cases this makes further exporting unnecessary, as it may be more practical to work directly with ELAN XML.

Basically, the built-in search functions in ELAN are similar to other on- or offline interfaces to written corpus data. However, one significant advantage of working with ELAN is that the same search functionalities of a local version of ELAN (see the description of ELAN above) can be utilized online – and thus off-site – to access all corpus files archived in the CMDI archive at the Max Planck Institute for Psycholinguistics in Nijmegen/Netherlands. For this, the tools ANNEX[22] and TROVA[23], which are basically online GUIs of the same search engines mentioned above for ELAN, and the metadata browser[24] at TLA can be used. ANNEX is an interface that links annotations and media files from the archive online (just as ELAN does on a local computer). The TROVA tool can be used to perform complex searches on multiple layers of the corpus and across multiple files in basically the same was as within ELAN itself. As a practical benefit, the integration of TROVA and whole infrastructure at TLA makes it easy to control corpus access, something that often demands significant consideration when sensitive language documentation materials are concerned. At the same time this infrastructure also allows examples to be disseminated more widely because it is easy to provide links for specific utterances in the data. Ultimately of course, the access rights of a specific file in the archive determine whether this works or not.

It should be noted that ELAN and other tools developed at the Max Planck Institute for Psycholinguistics are but one option for language documentation projects. There are other interoperable tools available for annotation, corpus management and corpus searches.[25] However, as the majority of these tools store data in XML format, importing data annotated with one tool into another tool requires only a standard XSL transformation.

For several reasons, the choice of Max Planck Institute for Psycholinguistics' platform for our data archiving project was a fortunate one. The platform offers a multitude of synergistically usable tools for storage and retrieving language data. Moreover, it makes use of the flexible metadata format CMDI, a replacement of the former IMDI[26] format. Since 2008, the CLARIN[27] initiative promoted CMDI for two main reasons:

1. the previous metadata formats proved to be either too superficial, such as OLAC[28]
2. or too specific, such as IMDI or TEI[29] (cf. Broeder and Uytvanck 2014).

[22] ANNEX – Annotation Explorer `https://tla.mpi.nl/tools/tla-tools/annex`

[23] TROVA – Search engine for annotation content archived at TLA `https://tla.mpi.nl/tools/tla-tools/trova`

[24] ARBIL – a general metadata editor, browser, and organizer `https://tla.mpi.nl/tools/tla-tools/arbil`

[25] Cf. for instance EXMARaLDA (Extensible Markup Language for Discourse Annotation) developed by the Hamburg Centre for Language Corpora and essentially featuring a similar functionality as ELAN `http://www.exmaralda.org/en/tool/exmaralda`.

[26] ISLE Meta Data Initiative `https://en.wikipedia.org/wiki/IMDI`

[27] `https://www.clarin.eu`

[28] Open Language Archives Community `http://www.language-archives.org`

[29] Text Encoding Initiative `http://www.tei-c.org`

Due to the fact that our data are not only in text format but also in audio or video format, and thus multimodal, CMDI is better suited for our purposes than the TEI metadata format. Last but not least, CMDI metadata files can be shared by many web services in a way that makes linguistic resources as accessible as possible to users. After all, by adopting the CMDI metadata format, our approach is perfectly in line with the endeavors undertaken by recent programs such as CLARIN and opens *digital humanities* specifically to marginalized Uralic minority speech communities.

4.3 Automated FST-based corpus annotation

Unlike many other endangered language documentation projects, which annotate spoken language data manually (or occasionally semi-manually, see above), we implement a significantly more automated method of corpus data annotation. Using the Giellatekno infrastructure (for more details, see Moshagen, Pirinen, et al. 2013; Johnson et al. 2013; Trosterud 2006b), we have started to implement FST-based language tools for our languages and to employ these in corpus annotation.

For morphological analysis, Giellatekno's infrastructure with the standard FST technology has been used for modeling stems, segmental affixes and other types of morphosyntactic information. The grammars are written within the *lexc* formalism and compiled with either the `lexc`[30] or the open source `hlexc`[31] compilers. The upper side of the resulting transducer consists of a lemma and a string of grammatical tags for each word form, while the lower side contains the concatenation of stem, affixes and markers signaling suprasegmental rules. The lower side of the *lexc* transducer is fed into *twolc*, a Two-Level-Morphology component (cf. Koskenniemi 1984; Moshagen, Trosterud, et al. 2008) used for handling complex suprasegmental morphological rules which are particularly characteristic of the Saamic languages we work with.

The process of annotation enrichment in ELAN is rather straightforward. The word forms of each utterance are extracted from the word-tier and sent to the morphosyntactic analyzer. The analysis output is then parsed and the bits of information are structured and put back into the ELAN file. Yet, as simple as it looks, the implementation required a careful analysis of item indexing in ELAN. On the one hand, all new annotation items have to land in the correct place in the structure, which involves keeping track of the respective indices for speaker, utterance, and word form. On the other hand, new indices have to be generated in such a way that they should not conflict with the extant indices assigned when an ELAN file is created. Since ELAN data can include the transcribed overlapping speech of several recorded speakers, it is not only three new tiers for `lemma`, `part-of-speech`, and `morphosyntactic description` that need to be generated and added to the initial structure, but 3xN, with N being the total number of speakers recorded in the ELAN file. If the new tiers were not generated and placed in the correct place, the ELAN XML structure would be spoiled, thus blocking the enriched ELAN file from showing up in the ELAN GUI as desired.

Now, let's have a closer look at the annotation process. Since the ELAN files are in XML format, they can be both read and edited by humans with any text editor and

[30] XFST tools `http://fsmbook.com`

[31] HFST tools `https://kitwiki.csc.fi/twiki/bin/view/KitWiki/HfstOverviewAndQuickStart`

accessed automatically by virtually any programming language. For the implementation of the script that links the ELAN data and the FST, we decided to use Python because:

1. it is a flexible, interpreted programming language with support for multiple systems and platforms;
2. it is easy to read and to learn for even a novice programmer, which is perhaps the reason why it is often used for linguistic applications;
3. and finally, it offers XML processing support by means of various XML packages such as ElementTree and lxml.

The input file for the whole process is an ELAN file lacking `lemma`, `part-of-speech`, and `morphological description` tiers. Thus, all tiers dependent on the `word`-tier (cf. Figure 1) are inserted dynamically. For each speaker recorded in the ELAN file, the values of each word form in each individual `word`-tier are extracted by the Python script and sent to the appropriate morphosyntactic analyzer. Since for each `word`-tier, the language is specified in the input data and the script takes this piece of information into account, it is possible to analyze each utterance with the FST of the specified language in a single run. This feature is especially useful for our data because most of our texts contain mixed language, mostly including both minority and majority language. This means that for ELAN files with several languages represented by the recorded speakers (for instance in the case of semi-communication between Kildin Saami and Ter Saami speakers in one and the same recording), the analysis is performed with the correct FST for each individual speaker (assuming this is specified correctly in the ELAN file).

After the FST has analyzed the word forms and has output the analyses, the Python script parses the FST output and restructures it when multiple analyses in a cohort are possible. A cohort is a word form along with all its possible analyses from the FST. An example of FST output is shown in Figure 2 in the upper left corner. Each individual lemma depends on the word form sent to the FST, each part-of-speech depends on a specific lemma, and finally each morphosyntactic description depends on a specific part-of-speech. With these constraints, new ELAN tiers for the analysis are built by factoring the different item types accordingly. For instance, the Pite Saami word form *vuolen* gets three different analyses, yet, with only two different lemmas *vuolen* and *vuolle*.

Ambiguities in language analyses are quite common, but with FSTs in development for minority languages, they are even more frequent. Our plan is to develop disambiguation modules based on how Giellatekno did this for Northern Saami: by using Constraint Grammar (cf. Antonsen et al. 2009). This takes morphologically analyzed text as its input and ideally only returns the appropriate reading, enriched with grammatical functions and dependency relations. Since the output of a CG is a dependency structure for a particular sentence, the output may also be converted into phrase structure representations. We plan to work on CGs for our languages and implement these in an extended FST/CG-ELAN script in the near future. However, despite well-developed disambiguation modules, when analyzed, even the Northern Saami corpus cannot be totally disambiguated.

Currently in our projects, we have decided to keep all ambiguities coming from the analysis in the ELAN file because there is an annotation mode in ELAN which enables users to correct annotation values in the data. At first glance, this can be interpreted as "semi-automatic" annotation, but it is more than that. The idea is that all corrections of a corpus analysis with an underdeveloped FST are fed back into the FST. This way the FST is constantly being improved and its coverage is extended incrementally. Unlike assigning morphosyntactic annotations from a simple 2-column table containing `word form–annotation`, using an FST is a laborious yet a far more viable, long-term solution. Aside from the fact that, while the FSTs for our working languages are already freely accessible, the creation of a 'simple' 2-column table with `word form–annotation` would be tantamount to having the data manually annotated, which is not the case.

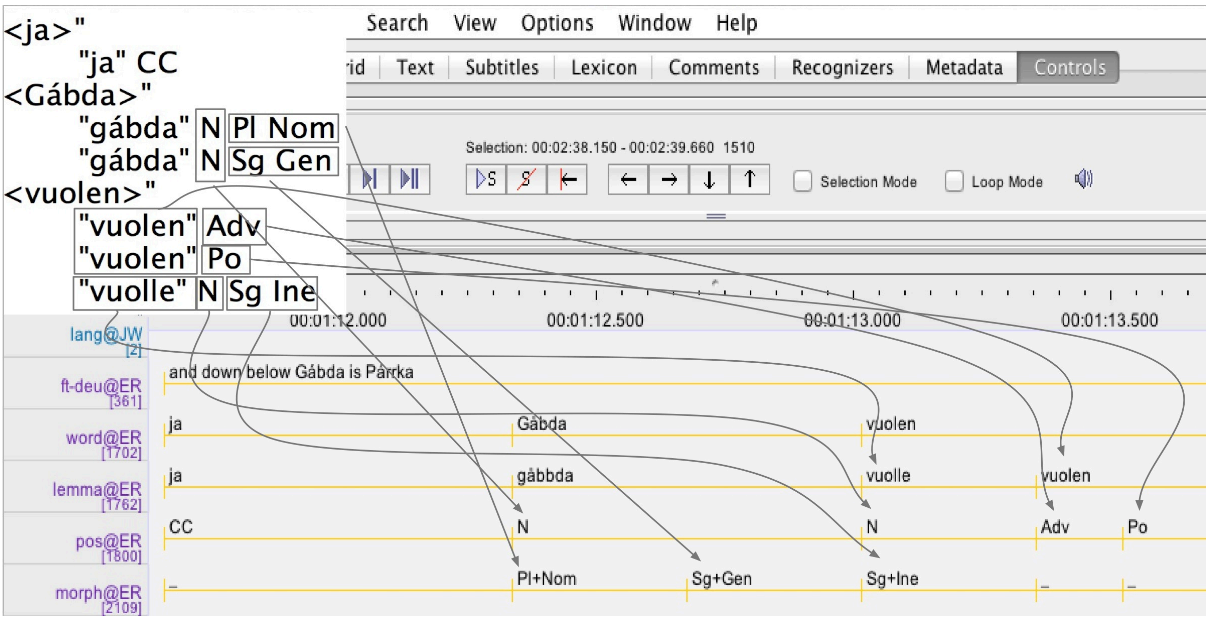
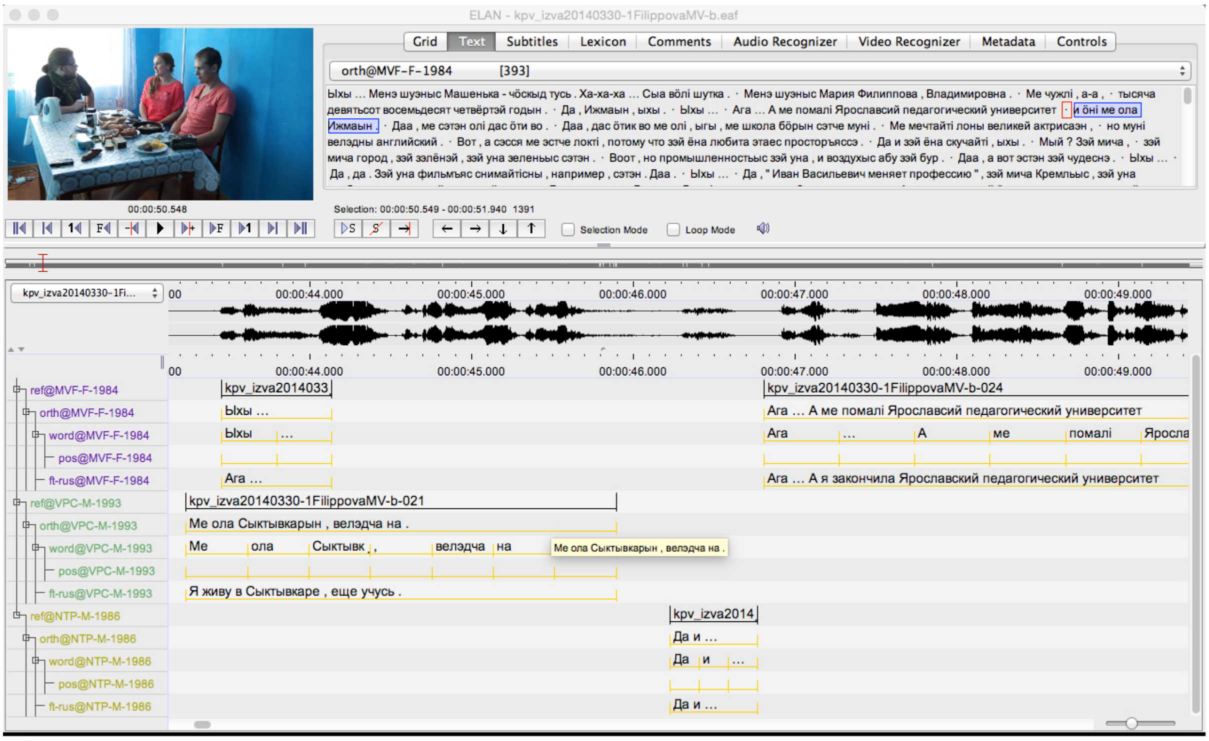


Figure 2: All morphosyntactic ambiguities of the FST are preserved in the ELAN structure, as shown in this Pite Saami example *ja Gábda vuolen* ("and down below Gábda")

So, by keeping all ambiguities in the analysis, the output of the morphosyntactic analysis is then segmented into `lemma`, `pos` and `morph` parts, transformed into the appropriate XML structure, and added to the ELAN file in new dependent tiers below the corresponding word form for each speaker recorded in the ELAN file (see Figure 2).

A further example worth illustrating concerns the annotation of multi-modal data. When we apply our script to an ELAN file, it first uses the built-in ELAN tokenizer to create a `word`-tier containing dependent annotations for each token encountered in the transcription found in the `orth`-tier. These tokens are then processed by the FST: when the Giellatekno FST is fed, for instance, with the string '`кулесьт`' 'at/from a fish' from Kildin Saami, it outputs '`кулль+N+Loc+Sg`', which consists of the relevant lemma '`кулл`', the part-of-speech category '`N`', and morphological information '`Loc`' and '`Sg`'. The resulting output is then teased apart into the `lemma`, `part-of-speech` and `morphology components`, and each of these is then turned into a dependent annotation in the new relevant ELAN `lemma`-tier, `pos`-tier and `morph`-tier, respectively. Figure 3 illustrates how ELAN presents an audio-video recording with annotations, including those created using FST.



orth@MVF-F-1984 [393]

Ыхы … Менэ шуэныс Машенька - чӧскыд тусь . Ха-ха-ха … Сыа вӧлі шутка . · Менэ шуэныс Мария Филиппова , Владимировна . · Ме чужлі , а-а , · тысяча девятьсот восемьдесят четвёртэй годын . · Да , Ижмаын , ыхы . · Ыхы … · Ага … А ме помалі Ярославсий педагогический университет · и ӧні ме ола Ижмаын . · Даа , ме сэтэн олі дас ӧти во . · Даа , дас ӧтик во ме олі , ыгы , ме школа бӧрын сэтче муні . · Ме мечтайті лоны великей актрисаэн , · но муні велэдны английский . · Вот , а сэсся ме эстче локті , потому что зэй ёна любита этаес просторъяссэ . · Да и зэй ёна скучайті , ыхы . · Мый ? Зэй мича , · зэй мича город , зэй зэлёнэй , зэй уна зеленьыс сэтэн . · Воот , но промышленностьыс зэй уна , и воздухыс абу зэй бур . · Даа , а вот эстэн зэй чудеснэ . · Ыхы … · Да , да . Зэй уна фильмъяс снимайтісны , например , сэтэн . Даа . · Ыхы … · Да , " Иван Васильевич меняет профессию " , зэй мича Кремльыс , зэй уна

Selection: 00:00:50.549 - 00:00:51.940 1391

kpv_izva2014033

Ыхы …

Ыхы …

Ага …

kpv_izva20140330-1FilippovaMV-b-024

Ага … А ме помалі Ярославсий педагогический университет

Ага … А ме помалі Яросла

Ага … А я закончила Ярославский педагогический университет

kpv_izva20140330-1FilippovaMV-b-021

Ме ола Сыктывкарын , велэдча на .

Ме ола Сыктывк , велэдча на

Ме ола Сыктывкарын , велэдча на .

Я живу в Сыктывкаре , еще учусь .

kpv_izva2014

Да и …

Да и …

Да и …

Figure 3: ELAN in player/annotation mode showing annotations, the audio waveform and the accompanying video for an Izhva Komi recording (session kpv_izva20140329-1 in Blokland, Fedina, et al. 2009–2017)

4.4 Conclusions and Prospects

We have shown that combining language documentation and language technology is a very promising undertaking for both fields, albeit for differing reasons. It is precisely in the overlapping areas between the two fields that a large amount of potential for the creation of resources useful in both fields can and should take place. Up to now, these complementary resources have hardly been utilized.

The simple yet effective model presented in this paper demonstrates how our language documentation projects take advantage of various tools of language technology. As a result of using our projects' corpora, which have both quantitatively and qualitatively superior annotations, language technology – in this case, Giellatekno – has access to new resources for further research. This is particularly the case concerning multimodal language corpora, which language technology and computer linguistics for Uralic languages have hardly dealt with up to now.

Our projects are still works in progress. There is no FST available for Akkala and Ter Saami, while the FST for Kildin Saami is still in an initial development phase. Pite Saami has already been developed further in this respect, particularly for nouns and verbs, but still need considerable work for other parts of speech. Tools for Skolt Saami and Komi, on the other hand have been under active development for some time, and already provide very good results.

Currently, we have only developed the script to automatically add automated FST-based morphological analyses (`lemma-pos-morph`) to ELAN annotations. As our projects continue, we will supplement and revise the FSTs for the languages described in this paper incrementally. Currently, FST technology can be readily implemented only for the languages for which models exist which are already in a rather mature state. While this is not an insignificant number of languages, the amount of work required to build well functioning analyzers should not be underestimated. The most advanced analyzers certainly are as good as they are simply due of the years of work spent developing with them. However, language documentation projects very often spend large amounts of resources on understanding the grammatical structures of the language in question and collecting large vocabularies. Maybe some ways could be found to turn these resources more directly into formats usable for FST-tools.

Note also that so far, our work with the script has resulted in a rather insular solution which is directly applicable to our own projects only; in order to make our workflow more usable for other projects we have to find a more generic solution. Potentially, our script could be integrated into the ELAN program using a web service. On the other hand, modifying the Python script to work with somewhat different ELAN input files would not be very difficult.

The corpus data that we plan to publish in the near future will also be available to interested parties in a variety of ways. On the one hand, ANNEX and TROVA can be used to browse and search the spoken corpora online at TLA. On a purely textual level – i.e., without alignment to multimedia – our corpora can also be integrated into other corpus interfaces for online browsing of written corpora, for example the Korp interface mentioned above, which is already in place for a number of languages at Giellatekno. Since some of the related multimedia is available online, one could also include links to the audio and video segments elsewhere in these more textual interfaces, even if this were not within the main functionality of the interface. would be An integrated dictionary with links to corpus data (such as Neahttadigisánit[32] – this already works very well for Northern Saami) would be another possible user interface that is particularly useful for language learners. However, this interface to corpus data is only textual, and does not have any links to multimedia recordings. On the other hand, providing access only to textual data has its own advantages, as it is possible to anonymize written transcriptions, whereas this may not be feasible with actual recordings. Since transcriptions are typically produced or at least evaluated by native speakers and are thus usually of rather high quality, even text versions alone are sufficient for many research purposes.

Acknowledgments

This paper is a revised and extended version of Blokland, Gerstenberger, et al. 2015. We would like to thank the other co-authors of the original paper, Rogier Blokland and Marina Fedina, for their ongoing collaboration. We also want to express our gratitude to Trond Trosterud and Jack Rueter for continuous valuable feedback. Thanks are also due to the two anonymous peer-reviewers for their useful comments. Last but not least, important ideas for our research were developed while writing the application for the

[32] `http://sanit.oahpa.no`

INEL project by Beáta Wagner-Nagy together with Michael Rießler and The Hamburg Center for Language Corpora (University of Hamburg). Sources of external funding for our research projects include the Kone Foundation, the German Research Foundation (DFG), and the Union of the German Academies of Sciences and Humanities.

References

Antonsen, Lene, S. Huhmarniemi, and Trond Trosterud (2009). "Constraint Grammar in dialogue systems". In: *NEALT Proceedings Series 2009.* Vol. 8. Tartu: Tartu ülikool, pp. 13–21.

Arkhangelskiy, Timofey and Maria Medvedeva (2016). "Developing morphologically annotated corpora for minority languages of Russia". In: *Proceedings of Corpus Linguistics Fest 2016. Bloomington, IN, USA, June 6–10, 2016.* Ed. by Sandra Kübler and Markus Dickinson. CEUR Workshop Proceedings 1607. Bloomington: Indiana University, pp. 1–6. URL: http://ceur-ws.org/Vol-1607/arckhangelskiy.pdf.

Austin, Peter K. (2013). "Language documentation and meta-documentation". In: *Keeping languages alive. Documentation, pedagogy and revitalisation.* Ed. by Mari Jones and Sarah Ogilvie. Cambridge: Cambridge University Press, pp. 3–15.

— (2014). "Language documentation in the 21st century". In: *JournaLIPP* 3, pp. 57–71. URL: http://lipp.ub.lmu.de/article/download/190/83.

Beáta, Wagner-Nagy and Sándor Szeverényi (2015). "Linguistically annotated spoken Nganasan corpus". In: *Tomsk Journal of Linguistics and Anthropology* 2, pp. 25–33.

Blokland, Rogier, Marina Fedina, Niko Partanen, and Michael Rießler (2009–2017). "Izhva Kyy". In: *The Language Archive (TLA). Donated Corpora.* In collab. with Vasilij Čuprov, Marija Fedina, Dorit Jackermeier, Elena Karvovskaya, Dmitrij Levčenko, and Kateryna AND Olyzko. Nijmegen: Max Planck Institute for Psycholinguistics. URL: https://corpus1.mpi.nl/ds/asv/?5&openhandle=hdl:1839/00-0000-0000-000C-1CF6-F.

Blokland, Rogier, Ciprian Gerstenberger, Marina Fedina, Niko Partanen, Michael Rießler, and Joshua Wilbur (2015). "Language documentation meets language technology". In: *First International Workshop on Computational Linguistics for Uralic Languages, 16th January, 2015, Tromsø, Norway. Proceedings of the workshop.* Ed. by Tommi A. Pirinen, Francis M. Tyers, and Trond Trosterud. Septentrio Conference Series 2015:2. Tromsø: The University Library of Tromsø, pp. 8–18. DOI: 10.7557/scs.2015.2.

Broeder, Daan and Dieter van Uytvanck (2014). "Metadata formats". In: *The Oxford handbook of corpus phonology.* Ed. by Jacques Durand, Ulrike Gut, and Gjert Kristoffersen. Oxford Handbooks. Oxford: Oxford University Press.

Comrie, Bernard, Andrey Shluinsky, and Olesya Khanina (2005–2017). "Documentation of Enets. Digitization and analysis of legacy field materials and fieldwork with last speakers". In: *Endangered Languages Archive (ELAR).* Digital language archive. London: SOAS University of London. URL: https://elar.soas.ac.uk/Collection/MPI950079.

Giellatekno and Divvun (2016). *SIKOR UiT The Arctic University of Norway and the Norwegian Saami Parliament's Saami text collection, Version 08.12.2016.* http://gtweb.uit.no/korp. Accessed: 2016-12-08.

Gippert, Jost, Ulrike Mosel, and Nikolaus Himmelmann, eds. (2006). *Essentials of language documentation.* Trends in Linguistics. Studies and Monographs 178. Berlin: Mouton de Gruyter.

Himmelmann, Nikolaus (2006). "Language documentation. What is it and what is it good for?" In: *Essentials of Language Documentation.* Ed. by Jost Gippert, Ulrike Mosel, and Nikolaus Himmelmann. Trends in Linguistics. Studies and Monographs 178. Berlin: Mouton de Gruyter, pp. 1–30.

— (2012). "Linguistic data types and the interface between language documentation and description". In: *Language Documentation & Conservation* 6, pp. 187–207. URL: `http://hdl.handle.net/10125/4503`.

Jauhiainen, Heidi, Tommi Jauhiainen, and Krister Lindén (2015). "The Finno-Ugric Languages and The Internet Project". In: *First International Workshop on Computational Linguistics for Uralic Languages, 16th January, 2015, Tromsø, Norway. Proceedings of the workshop.* Ed. by Tommi A. Pirinen, Francis M. Tyers, and Trond Trosterud. Septentrio Conference Series 2015:2. Tromsø: The University Library of Tromsø, pp. 87–98. DOI: `10.7557/5.3471`.

Johnson, Ryan, Lene Antonsen, and Trond Trosterud (2013). "Using finite state transducers for making efficient reading comprehension dictionaries". In: *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), May 22–24, 2013, Oslo.* Ed. by Stephan Oepen and Janne Bondi Johannessen. Linköping Electronic Conference Proceedings 85. Linköping: Linköping University, pp. 59–71. URL: `http://emmtee.net/oe/nodalida13/conference/45.pdf`.

Koskenniemi, Kimmo (1984). "A General Computational Model for Word-form Recognition and Production". In: *Proceedings of the 10th International Conference on Computational Linguistics and 22Nd Annual Meeting on Association for Computational Linguistics.* ACL '84. Stanford, California: Association for Computational Linguistics, pp. 178–181. DOI: `10.3115/980491.980529`. URL: `http://dx.doi.org/10.3115/980491.980529`.

Lagercrantz, Eliel (1957–1966). *Lappische Volksdichtung.* 7 vols. Suomalais-ugrilaisen Seuran toimituksia 112,115,117,120,124,126,141. Helsinki: Suomalais-Ugrilainen Seura.

Moshagen, Sjur, Tommi A. Pirinen, and Trond Trosterud (2013). "Building an open-source development infrastructure for language technology projects". In: *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), May 22–24, 2013, Oslo.* Ed. by Stephan Oepen and Janne Bondi Johannessen. Linköping Electronic Conference Proceedings 85. Linköping: Linköping University, pp. 343–352. URL: `http://emmtee.net/oe/nodalida13/conference/43.pdf`.

Moshagen, Sjur, Trond Trosterud, and Pekka Sammallahti (2008). "Twol at work". In: *Inquiries into Words, Constraints and Contexts.* Ed. by Antti Arppe, Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna Westerlund, and Anssi Yli-Jyrä. Stanford: CSLI, pp. 94–105.

Nagy, Naomi and Miriam Meyerhoff (2015). "Extending ELAN into variationist sociolinguistics". In: *Linguistics Vanguard* 1.1, pp. 271–281. DOI: `DOI10.1515/lingvan-2015-001`.

Partanen, Niko, Alexandra Kellner, Timo Rantakaulio, Galina Misharina, and Hamel Tristan (2013). "Down River Vashka. Corpus of the Udora dialect of Komi-Zyrian". In: *The Language Archive (TLA). Donated Corpora.* Nijmegen: Max Planck Institute for Psycholinguistics. URL: `https://hdl.handle.net/1839/00-0000-0000-001C-D649-8`.

Poibeau, Thierry and Benjamin Fagard (2016). "Exploring natural language processing methods for Finno-Ugric languages". In: *Second International Workshop on Computational Linguistics for Uralic Languages, 20th January, 2016, Szeged, Hungary. Proceedings of the workshop.* Ed. by Tommi A. Pirinen, Francis M. Tyers, and Trond Trosterud. Szeged: University of Szeged. In press.

Rießler, Michael (2005–2017). "Kola Saami Documentation Project. Linguistic and ethnographic documentation of the endangered Kola Saami languages". In: *The Language Archive (TLA). DoBeS archive.* Digital language archive. In collab. with Anna Afanas'eva, Anja Behnke, Svetlana Danilova, Andrej Dubovcev, Aleksandra Erštadt, Dorit Jackermeier, Elena Karvovskaya, Kristina Kotcheva, Jurij Kusmenko, Maryna Litvak, Sergej Nikolaev, Kateryna Olyzko, Niko Partanen, Elisabeth Scheller, Nina Šarshina, Ganna Vinogradova, Joshua Wilbur, Evgenia Zhivotova, and Nadežda Zolotuchina. Nijmegen: Max Planck Institute for Psycholinguistics. URL: `https://corpus1.mpi.nl/ds/asv/?2&openhandle=hdl:1839/00-0000-0000-0005-8A34-E`.

Snoek, Conor, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud (2014). "Modeling the noun morphology of Plains Cree". In: *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages.* Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 34–42. URL: `http://www.aclweb.org/anthology/W/W14/W14-2205`.

Trosterud, Trond (2006a). "Grammar-based language technology for the Sámi Languages". In: *Lesser used Languages & Computer Linguistics.* Bozen: Europäische Akademie, pp. 133–148.

— (2006b). "Grammatically based language technology for minority languages. Status and policies, casestudies and applications of information technology". In: *Lesser-known languages of South Asia.* Ed. by Anju Saxena and Lars Borin. Trends in Linguistics. Studies and Monographs 175. Berlin: Mouton de Gruyter, pp. 293–316.

Wilbur, Joshua (2008–2017). "Pite Saami. Documenting the language and culture". In: *Endangered Languages Archive (ELAR).* Digital language archive. In collab. with Iris Perkmann, Elsy Rankvist, and Peter Steggo. London: SOAS University of London. URL: `https://elar.soas.ac.uk/Collection/MPI201072`.

— (2011). "Think Globally, Archive Locally. Opportunities and challenges in working with local archiving institutions". In: *Proceedings of the Workshop on Language Documentation and Archiving.* Ed. by David Nathan. London: SOAS University of London, pp. 51–58.

— (2014). "Archiving for the community. Engaging local archives in language documentation projects". In: *Language Documentation and Description* 12: *Special Issue on Language Documentation and Archiving.* Ed. by David Nathan and Peter K. Austin, pp. 85–102. URL: `http://www.elpublishing.org/PID/139`.

Woodbury, Anthony C. (2011). "Language documentation". In: *The Cambridge handbook of endangered languages.* Ed. by Peter K. Austin and Julia Sallabank. Cambridge handbooks in language and linguistics. Cambridge: Cambridge University Press, pp. 159–186.