Proceedings of 1st International Workshop in Computational Linguistics for Uralic Languages (IWCLUL 2015); ‹http://dx.doi.org/10.7557/scs.2015.2›

Language Documentation meets Language Technology

Rogier Blokland
Uppsala universitet (Institutionen för moderna språk)
rogier.blokland@moderna.uu.se

Marina Fedina
Komi Republican Academy of State Service and Administration
(Centre for Innovative Language Technology)
fedinamarina@gmail.com

Ciprian Gerstenberger
UiT Norges arktiske universitet (Giellatekno)
ciprian.gerstenberger@uit.no

Niko Partanen
Albert-Ludwigs-Universität Freiburg (Skandinavisches Seminar)
niko.partanen@skandinavistik.uni-freiburg.de

Michael Rießler
Albert-Ludwigs-Universität Freiburg (Skandinavisches Seminar)
michael.riessler@skandinavistik.uni-freiburg.de

Joshua Wilbur
Albert-Ludwigs-Universität Freiburg (Skandinavisches Seminar)
joshua.wilbur@skandinavistik.uni-freiburg.de

December 16, 2014

Abstract

The paper describes work-in-progress by the Pite Saami, Kola Saami and Izhva Komi language documentation projects, all of which use similar data and technical frameworks and are carried out collaboratively in Uppsala, Tromsø, Syktyvkar and Freiburg. Our projects record and annotate spoken language data in order to provide comprehensive speech corpora as databases for future research *on* and *for* these 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. Ultimately, the multimodal corpora created by our projects will be useful for scientifically significant quantitative investigations on these languages in the future.

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. 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 spoken language data. Metadata annotations are crucial for the intellectual accessibility of the documented data and concern both the *content* of the recorded speech sample (typically represented as phonological, morphological or syntactic transcriptions and translations) as well as the *context* (such as actors, places, speech events, but also meta-documentation about the actual project).

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 over the last years. In fact, with respect to the data types involved, documentary linguistics generally seems similar to corpus building in principle. Both

provide primary data for secondary (synchronic or diachronic) data derivations and analyses. The main difference is that traditional corpus and computational linguistics deal 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, results in rather small corpora of exclusively spoken genres. Furthermore, corpus annotations in language documentation projects are often created manually. Significant quantitative investigations based on corpora from language documentation projects are therefore normally excluded.

Language documentation has made huge technological progress in regard to collaborative tools and user interfaces for transcribing, archiving and browsing multimedia recordings. However, paradoxically, the field has only rarely considered applying automated methods to more efficiently (both qualitatively and quantitatively) annotate data in creating a basis for new and better corpus-based linguistic research on smaller languages.

Although the relevant methods and tools would be completely functional even for relatively small languages such as North Saami today, they are being applied exclusively for corpus-building of *written* language varieties. Current language technology projects on endangered languages (e.g. Giellatekno[1]) 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 such minority languages, but represent a rather limited range of text genres. Furthermore, as the current written standards of small endangered languages (e.g. North Saami) are to a large part evolving as the result of institutional language planning, the bulk of texts in the North Saami corpus consists of translations from the majority languages, and even original Saami texts (e.g. on official webpages and in the few newspapers) are most typically produced only by a few writers.

The restriction on written language is even more crucial in the case of smaller languages such as Skolt Saami, for which language technology is also under development. 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 used in speech only. As a consequence, there is a need to enrich the existing corpora for languages such as North 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 non-written modi, and an efficient and consistent method for incorporating spoken texts is vital for corpus creation. In fact, spoken language documentations for these languages exist and several projects continue collecting new and annotating legacy

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

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 met so far.

Our projects are concerned with the building of multimodal corpora (at least, spoken and written, i.e., transcribed), and thus form an interface between endangered language documentation and 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 are but two examples of such practical applications. We hope to show 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 analyzes necessary for the computational modeling of the languages in question, language technologists apply formal-descriptive linguistic and corpus linguistic methods to the programming of machine-readable grammatical and lexical descriptions of the relevant languages. Spoken language documentations can thus increase the size of the data pool utilized in computational linguistic research. Language technology, on the other hand, can create tools for effectively analyzing spoken language corpora and carrying out better linguistic documentation and description on the endangered languages in question.

2 Language Documentation meets Language Technology

This paper describes our current work on recording and annotation spoken language data and discusses the combined methods from language documentation and language technology used by our projects. The languages we are working on at present are Pite Saami, Skolt Saami, Kildin Saami[2] and the Izhva variety of Komi-Zyrian.[3] Illustrated with data examples from our current projects we will show how *language documentation* profits from the application of automated corpus data annotation, specifically Finite State Transducer technology (hereinafter FST), which not only helps provide (quantitatively and qualitatively) enhanced annotations, but ultimately results in better databases useful for (quantitative and qualitative) corpus-linguistic research. *Language technology*, on the other hand, can profit from from the use of more extensive

[2] http://www.skandinavistik.uni-freiburg.de/forschung/forschungsprojekte/saami
[3] http://komikyv.ru/page/about

and more diverse data.

In addition to designing annotation schemata of appropriate granularity for corpus building, two essential aspects of language documentation remain important for our own approach: the *archiving* of primary data linked to all data derivations as well as proper *contextualization* by means of deep metadata. By 'deep medatada' we mean metadata concerning a variety of levels of description in addition to basic cataloguing 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, which are nevertheless crucial for language documentation aiming at long lasting comprehensive, multi-faceted and multipurpose records of linguistic practices. It is also worth mentioning that our approach is perfectly in line with the endeavors made by recent programs such as CLARIN[4] and opens *digital humanities* for marginalized Uralic minority speech communities specifically.

3 ELAN as a tool for annotating multimodal corpora

The language corpora we are building represent spoken and written text modi of formal and informal registers and a variety of genres. Our transcribed (in standard orthography) spoken text data as well as the written text data are stored in XML format and structured to be utilized by the multimedia annotation program ELAN.[5] This software allows audio and video recordings to be time aligned with detailed, multi-level transcriptions, translations and further annotations. Furthermore, with ELAN basic frequency statistics can be calculated, concordances created, and data for statistical analysis exported (e.g., using R[6] or similar tools).

Annotation tiers in the ELAN files from our projects are organized hierarchically based on the minimal template in Figure 1 for each speaking participant in a recording. Since each speaker has his/her own tier node `ref`, including dependent tiers, annotating simultaneous speech by multiple speakers (a common feature of spoken language) is not problematic.

While ELAN is intended mainly as an interface between written transcriptions/annotations and the original audio/video medium in which every annotation is time aligned with the medium, it is also possible to use ELAN for texts in written form and without audio/video. In this way, written legacy texts are also included in the corpora

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

[5] ELAN is free software developed by the Technical Group of the Max Planck Institute for Psycholinguistics, see `https://tla.mpi.nl/tools/tla-tools/elan`.

[6] `http://www.r-project.org`

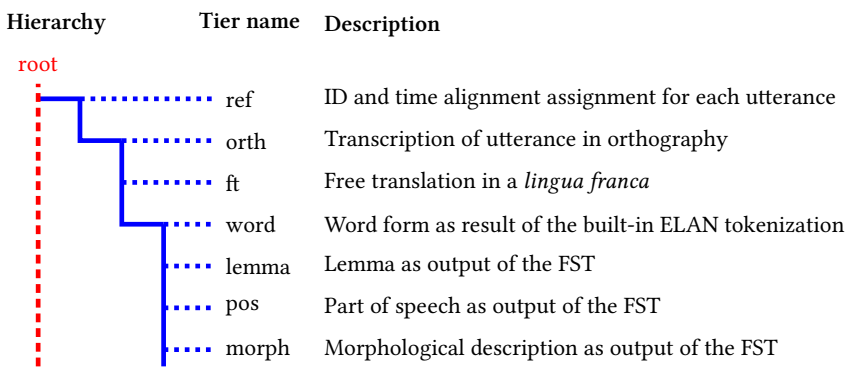

Figure 1: The basic ELAN tier hierarchy used in the documentation corpora

we create.

Metadata stored in IMDI format[7] can also be 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 instruments used, as well as about work with specific projects or persons can be included into metadata . It is also desirable to store metadata separated from basic annotations, as this makes it very easy to control access to more sensitive pieces of information that might be stored in the metadata. In our model, individual session names and actor IDs can be used to associate any metadata with any transcription.

We are already able to carry out corpus-wide searches on instances of a specific genre by using constraints, for instance, on participants' ages or regional affiliations. This provides a solid fundament for more fine-grained sociolinguistically oriented research. In principle, the transcription files also contain small traces of metadata, as the filenames themselves are standardized to include the language ISO-code and the recording date. Yet, with this data alone it is not possible to filter results with more contextual factors. Such filtering is only possible using with the associated metadata. Furthermore, it is possible to execute complex searches on multiple ELAN files – on the entire corpus or only specific parts of it. In this, search constraints on the type of tier, contextual information, etc. and regular expressions can be specified.

Search results can be shown in a *key-words in context* (KWIC) format, i.e. in a concordance where up to eight words on either side of the search term are visible. As for exporting, all search results can be saved in plain-text comma-separated-value

[7]For the `ISLE Meta Data Initiative` format see `http://www.mpi.nl/imdi`.

(csv) format. Finally, ELAN files are plain text files in XML structure (with the file extension .eaf), and as such are archive-friendly, somewhat human-readable, and will likely be supported well into the future as XML is a common and open-source format.

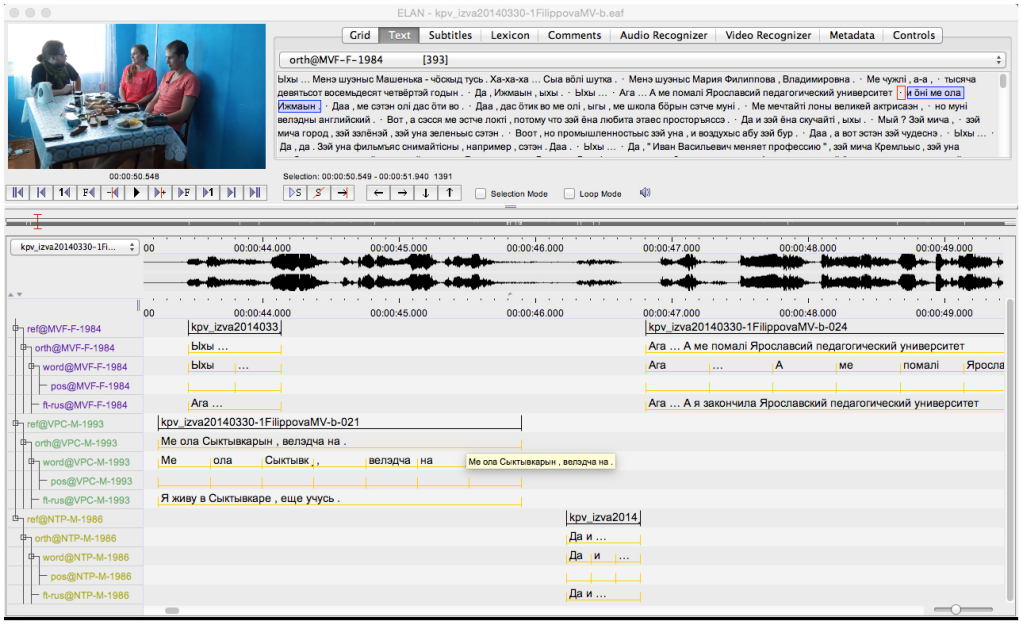

Figure 2: ELAN in player/annotation mode showing annotations for the overlapping speech of three speakers, the audio waveform, the accompanying video

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 IMDI archive at the Max Planck Institute for Psycholinguistics in Nijmegen/Netherlands. For this, the tool ANNEX[8] is an interface that links annotations and media files from the archive online (just like ELAN on a local computer). The TROVA tool[9] can be used to perform complex searches on multi-layers of the corpus and across multiple files.

[8]https://tla.mpi.nl/tools/tla-tools/annex
[9]https://tla.mpi.nl/tools/tla-tools/trova

4 Automated FST-based corpus annotation

Unlike many other endangered language documentation projects, which annotate spoken language data manually – or occasionally semi-manually – we apply a more automated way of corpus data annotation. Using the Giellatekno infrastructure, we have started by implementing FST-based language tools for Kildin Saami and employing these for corpus annotation. Since the ELAN files are XML files, they can be accessed by virtually any programming language with XML-processing support.

The process of annotation enrichment is quite simple. The input file for the whole process is an ELAN file without part of speech, lemma, or morphological description tiers. A Python script accesses the input file, takes each item from the word form tier and passes this to the morphosyntactic analyzer. The result is then segmented into `lemma`, `pos` and `morph` parts, transformed into the appropriate XML structure, and then loaded back into the input file.

As it is still in an initial development phase, Kildin Saami lacks language analysis tools on higher levels such as a disambiguation or a parsing module, usually implemented by means of a Constraint Grammar. This means that the result can consist of multiple analyses for the same word form. Since the analyses are split by `lemma`, `pos`, and `morph`, one might think that decoupling them and putting them on different levels (see Fig. 1) would lead to even more ambiguity. For instance, the `morph` category Perf does not fit the pos N. Yet, this is only superficially the case, internally, ELAN has a good pointing system between the tiers, hence, it is possible to point from, for instance, the `morph` annotation Comp to a pos annotation A. That way, the pieces of information coming from the FST are guaranteed to be places and linked properly.

This method is also beneficial for the further development of language analysis data. As mentioned above, the resources for Kildin Saami are still in an initial phase, and therefore the FST does not produce an analysis for some word forms. In such cases, the word form under scrutiny would be assigned a specific value for non-extant results. These can then be corrected manually by means of the ELAN tool and the improvements would then flow back into the FST resources. Subsequently running a corpus analysis would then produce better results.

5 Conclusion and prospects

We hope to 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 example 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 corpora, which language technology and computer linguistics for Uralic languages have hardly dealt with up to now.

Our projects are a work in progress. Currently, we have only developed a part-of-speech tagger for Kildin Saami. At the next stage, we intend to have complete morphological analyses (`lemma-pos-morph`) created automatically. Analyzing the corpus with help of FSTs on the morphosyntactic level can sometimes lead to cases of ambiguity. For disambiguation and syntactic analysis, Giellatekno uses *Constraint Grammar* (CG), which takes morphologically analyzed text as input, and ideally returns only 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 implement the infrastructure that we now are building for Kildin Saami for other languages for which FST already exists as well.[10] As our projects are carried out, we will continue to supplement and revise the FSTs for these languages incrementally.

The corpus data that we will archive in the near future shall 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 multimodal corpora online. On a purely textual level – i.e., without links to multimedia – our corpora can also be integrated into the Korp interface (a tool for online browsing of written corpora[11]), which is already in place for a number of languages at Giellatekno. Another possible user interface, which is particularly useful for language users, is an integrated dictionary with links to corpus data, such as Neahttadigisánit[12] – this already works very well for North Saami. However, this interface is only textual, and does not have any links to multimeda recordings.

[10]Cf. the respective documentation at Giellatekno.

[11]`http://gtweb.uit.no/korp`

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

References

[1] L. Antonsen, S. Huhmarniemi, and T. Trosterud. Constraint grammar in dialogue systems. In *Nealt proceedings series 2009*. Volume Volume 8, 2009, pages 13–21.

[2] Peter K. Austin. Language documentation and meta-documentation. In Mari Jones and Sarah Ogilvie, editors, *Keeping languages alive. Documentation, pedagogy and revitalisation*, pages 3–15. Cambridge University Press, Cambridge, 2013.

[3] Peter K. Austin. Language documentation in the 21st century. *JournaLIPP*, 3:57–71, 2014.

[4] Rogier Blokland, Michael Rießler, Niko Partanen, Marina Fedina, and Andrej A. Čemyšev. Ispol'zovanie cifrovych korpusov i komp'juternych program v dialektologičeskich issledovanijach. Teorija i praktika. In F. G. Chisamitdinova, editor, *Aktual'nye problemy dialektologii jazykov narodov rossii. Materialy xiv vserossijskoj naučnoj konferencii (ufa, 20–22 nojabrja 2014 g.)* Pages 252–255. IIJaL UNC RAN, Ufa, 2014.

[5] Nikolaus P. Himmelmann. Documentary and descriptive linguistics. *Linguistics*, 36:161–195, 1998.

[6] Nikolaus P. Himmelmann. Linguistic data types and the interface between language documentation and description. *Language documentation & conservation*, 6:187–207, 2012. URL: `http://hdl.handle.net/10125/4503`.

[7] Ryan Johnson, Lene Antonsen, and Trond Trosterud. 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*, number 16 in NEALT Proceedings Series, pages 59–71. Oslo University, Oslo, 2013. URL: `http://www.ep.liu.se/ecp_article/index.en.aspx?issue=085;article=010`.

[8] Sjur Moshagen, Trond Trosterud, and Pekka Sammallahti. Twol at work. In Antti Arppe, Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna Westerlund, and Anssi Yli-Jyrä, editors, *Inquiries into words, constraints and contexts*, pages 94–105. CSLI, Stanford, 2008.

[9] Sjur Moshagen, Tommi A. Pirinen, and Trond Trosterud. Building an open-source development infrastructure for language technology projects. In *Proceedings of the 19th Nordic Conference of Computational Linguistics NODALIDA*. (16) in NEALT Proceedings Series, 2013.

[10] Naomi Nagy and Miriam Meyerhoff. Extending ELAN into quantitative sociolinguistics. ePoster presented at ICLDC 3, University of Hawai'i at Mānoa, 28 February–3 March 2013. 2013.

[11] Trond Trosterud. Grammatically based language technology for minority languages. Status and policies, casestudies and applications of information technology. In Anju Saxena and Lars Borin, editors, *Lesser-known languages of South Asia*, number 175 in Trends in Linguistics. Studies and Monographs, pages 293–316. Mouton de Gruyter, Berlin, 2006.

[12] Anthony C. Woodbury. Language documentation. In Peter K. Austin and Julia Sallabank, editors, *The Cambridge handbook of endangered languages*, pages 159–186. Cambridge University Press, Cambridge, 2011.

[13] Anthony C. Woodbury. Archives and audiences. Toward making endangered language documentations people can read, use, understand, and admire. In David Nathan and Peter K. Austin, editors, *Language documentation and description*. Volume 12, pages 19–36. SOAS, London, 2014.

[14] Michael Rießler. Towards a digital infrastructure for Kildin Saami. In Erich Kasten and Tjeerd de Graaf, editors, *Sustaining indigenous knowledge. Learning tools and community initiatives on preserving endangered languages and local cultural heritage*, in SEC Publications. Exhibitions & Symposia series, pages 195–218. Verlag der Kulturstiftung Sibirien, Fürstenberg, 2013. URL: `http://www.siberian-studies.org/publications/PDF/sikriessler.pdf`.

[15] Trond Trosterud and Marina S. Fedina. Rol' jazykovoj technologii v sochranenii i revitalizacii jazyka. In T. V. Juzykajn, Andrej V. Šemyšev, and Ė. A. Juzykajn, editors, *Jazyki men'šinstv v kom'pjuternych technologijach. Opyt, zadači i perspektivy*. Materialy meždunarodnoj konferencii, pages 36–45. Joškar-Ola, 2011.

[16] Joshua Wilbur. Archiving for the community. Engaging local archives in language documentation projects. In David Nathan and Peter K. Austin, editors, *Language documentation and description*, pages 85–102. SOAS, London, 2014.