Instant annotations in ELAN corpora of spoken and written Komi, an endangered language of the Barents Sea region

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
Freiburg Institute for Advanced Studies
michael.riessler@frias.uni-freiburg.de

Abstract

The paper describes work-in-progress by the Izhva Komi language documentation project, which records new spoken language data, digitizes available recordings and annotate these multimedia data in order to provide a comprehensive language corpus as a databases for future research *on* and *for* this endangered – and under-described – Uralic speech community. While working with a spoken variety and in the framework of *documentary linguistics*, we apply *language technology* methods and tools, which have been applied so far only to normalized written languages. Specifically, we describe a script providing interactivity between ELAN, a Graphical User Interface tool for annotating and presenting multimodal corpora, and different morphosyntactic analysis modules implemented as Finite State Transducers and Constraint Grammar for rule-based morphosyntactic tagging and disambiguation. Our aim is to challenge current manual approaches in the annotation of language documentation corpora.

1 Introduction

Endangered language documentation (aka *documentary linguistics*) has made huge technological progress in regard to collaborative tools and user interfaces for transcribing, searching, and archiving multimedia recordings. However, paradoxically, the field has only rarely considered applying NLP methods to more efficiently annotate qualitatively and quantitatively language data for endangered languages, and this, despite the fact that the relevant computational methods and tools are well-known from corpus-driven linguistic research on larger written languages. With respect to the data types involved, endangered language documentation generally seems similar to corpus linguistics (i.e. "corpus building") of non-endangered languages. Both provide primary data for secondary (synchronic or diachronic) data derivations and analyses (for data types in language documentation, cf. Himmelmann 2012; for a comparison between corpus linguistics and language documentation, cf. Cox 2011).

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 spoken genres.

Although relatively small endangered languages are also increasingly gaining attention by the computational linguistic research (as an example for Northern Saami, see Trosterud 2006a; and for Plains Cree, see Snoek et al. 2014), these projects work predominantly with *written* language varieties. Current computational linguistic projects on endangered languages seem to have simply copied their approach from already established research on the major languages, including the focus on written language. The resulting corpora are impressively large and include higher-level morphosyntactic annotations. However, they represent a rather limited range of text genres and include predominantly translations from the relevant

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

majority languages.

In turn, researchers working within 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 broad variety of genres. 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),[1] FieldWorks Language Explorer (or FLEx, for short),[2] 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 or video recordings, the resulting corpora are relatively small and rarely reach 150,000 word tokens. Typically, they are considerably smaller. The main reason for the limited size of such annotated language documentation corpora is that manual glossing is an extremely time consuming task and even semi-manual glossing using FLEx or similar tools is a bottleneck because disambiguation of homonyms has to be solved manually.

Another problem we identified especially in the documentation of endangered languages in Northern Eurasia is that sometimes the existence of orthographies is ignored, and instead, phonemic or even detailed phonetic transcription is employed. It is a matter of fact that as a result of institutionalized and/or community-driven language planning and revitalization efforts many endangered languages in Russia have established written standards and do regularly use their languages in writing. Russia seems to provide a special case compared to many other small endangered languages in the world, but for some of these languages, not only Komi-Zyrian, but also for instance Northern Khanty, Northern Selkup, or Tundra Nenets (which are all object to documentation at present), a significant amount of printed texts can be found in books and newspapers printed during different periods, and several of these languages are also used digitally on the Internet today.[3] Compared to common practice in corpus building for non-endangered languages or dialects (for Russian dialects, cf. Waldenfels et al. 2014), the use of phonemic transcriptions in the mentioned language documentation projects seems highly anachronistic if orthographic systems are available and in use. Note also, that many contemporary and historical printed texts written in these languages are also already digitized,[4] which would make it easily possible in principle to combine spoken and written data in one and the same corpus later. The use of an orthography-based transcription system not only ease, and hence, speeds up the transcription process, but it enables a straightforward integration of all available (digitized) printed texts into our corpus in a uniform way, too. Note also, that combining all available spoken and written data into one corpus would not only make the corpora larger token-wise, but such corpora will also be ideal for future corpus-based variational sociolinguistic investigations. Falling back to orthographic transcriptions of our spoken language documentation corpora seems therefore most sensible. Last but not least, if research in language documentation is intended to be useful for the communities as well, the representation of transcribed texts in orthography is the most logical choice.

In our own language documentation projects, we question especially the special value given to time consuming phonemic transcriptions and (semi-)manual morpheme-by-morpheme interlinearizations when basic phonological and morphological descriptions are already available (which is arguably true for the majority of languages in Northern Eurasia), as such descriptions serve as a resource to accessing phonological and morphological structures. Instead, we propose a step-by-step approach to reach higher-level morphosyntactic annotations by using and incrementally improving genuine computational methods. This procedure encompasses a systematic integration of all available resources into the language documentation endeavor: textual, lexicographic, and grammatical.

The language documentation projects we work with (cf. Blokland, Gerstenberger, et al. 2015; Gerstenberger et al. 2017b; Gerstenberger et al. 2017a) are concerned with the building of mul-

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

[2] `http://fieldworks.sil.org/flex`

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

[4] 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, cf. `http://komikyv.ru`.

timodal 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. Language technology 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. A morphosyntactic tagger applied in corpus building is but one example of such a practical application. Using automated tagging, rather than the non-automated methods described in the previous section, allows for directing more resources towards transcription and for working with larger data sets because slow manual annotation no longer necessarily forms a bottleneck in a project's data management workflow.

The examples in our paper are taken specifically from Komi-Zyrian, although we work with several endangered Uralic languages spoken in the Barents Sea Area. We are developing a common framework for these different language projects in order to systematically apply methods from Natural Language Processing (NLP) to enrich the respective corpora with linguistic annotations. In order to do so, we rely on the following two main principles, which we have begun implementing consistently in our own documentation projects:

1. the use an *orthography-based transcription* system, and
2. the application of *computer-based methods* as much as possible in *creating higher-level annotations* of the compiled corpus data.

2 Available Data and Tools for Dialectal and Standard Komi

Izhva Komi is a dialect of Komi-Zyrian, henceforth Komi, spoken in the Komi Republic as well as outside the republic in several language islands in Northeastern Europe and Western Siberia. Beside Izhva dialect data we have also a smaller amount of data from the Udora dialect, spoken in the West of the Komi Republic. Komi belongs to the Permic branch of the Uralic language family and is spoken by approximately 160,000 people who live predominantly in the Komi Republic of the Russian Federation. Although formally recognized as the second official language of the Republic, the language is endangered as the result of rapid language shift to Russian.

Komi is a morphologically rich and agglutinating language with primarily suffixing morphology and predominantly head-final constituent order, differential object marking, and accusative-nominative alignment. The linguistic structure of Komi and its dialects is relatively well described (for an English language overview, cf. Hausenberg 1998). However, the existing descriptions focus on phonology and morphology and are not written in modern descriptive frameworks.

The most important official actor for (corpus and status) language planning for standard written Komi is the Centre for Innovative Language Technology at the Komi Republican Academy of State Service and Administration in Syktyvkar, which is currently creating the Komi National Corpus, a corpus that already now contains over 30M words. The spoken data we work with come from our own "Iźva-Komi Documentation Project" (2014–2016) and include fully transcribed and translated dialect recordings on audio and (partly) video. These recordings were mostly collected during fieldwork or origin from legacy data. The Udora data are similar in structure and origin from the project "Down River Vashka" (2013). An overview of the Komi data we have at our disposal is shown in Table 1.

Our language documentations are archived at The Language Archive (TLA, for short) at the Max Planck Institute for Psycholinguistics in Nijmegen/Netherlands (cf. Partanen et al. 2013; Blokland, Fedina, et al. 2009–2017). For the written Komi data, see Fedina et al. (2016–2017) and the online portal to the above mentioned Centre's data and tools, which is called "The Finno-Ugric Laboratory for Support of the Electronic Representation of Regional Languages".[5] The Centre has also made free electronic dictionaries and a Hunspell checker (including morpheme lists) available, but research towards a higher-level grammar parser has not been carried out in Syktyvkar so far.

Another important open-source language technology infrastructure for Komi is under development by Jack Rueter (Helsinki) at *Giella-*

[5] `http://fu-lab.ru`

Table 1: Overview on the amount of Komi spoken and written 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; note also that typically, our data include translations into at least one majority language too.

Language		**Modality**	**Recorded speakers/writers**	**Time span of texts**	**Tokens in corpus**
Komi-Zyrian	(Standard)	written	∼2,500	1920–2017	30,000,000
Komi-Zyrian	(Izhva dialect)	spoken	∼150	1844–2016	200,000
Komi-Zyrian	(Udora dialect)	spoken	∼50	1902–2013	40,000

tekno/Divvun – Saami Language Technology at UiT The Arctic University of Norway.[6] The Giellatekno group works with computational linguistic research into the analysis of Saamic and other languages of the circumpolar area. Giellatekno has the know-how and the infrastructure necessary to deal with all aspects of corpus and computational linguistics and has fully implemented this for Northern Saami (cf. Moshagen et al. 2013; Johnson et al. 2013; Trosterud 2006b).

The project described in this paper makes use of the infrastructure and tools already available in Syktyvkar and Tromsø, but works specifically with corpus construction and corpus annotation of spoken data, which have not been in focus of computational linguistic research so far.

3 Data Annotation Process

Nowadays, there is a multitude of approaches for Natural Language Processing (NLP) with 'pure' statistic-based on one end of a scale, 'pure' rule-based on the other end, and a wide range of hybridization in between (cf. also a recent "hybrid" approach using a manually interlinearized/glossed language documentation corpus from Ingush as training data for a tagger, Tiedemann et al. 2016).

For major languages such as English, Spanish, or Russian, the dominating paradigm within computational linguistics is based on statistical methods: computer programs are trained to understand the behavior of natural language by means of presenting them with vast amounts of either unanalyzed or manually analyzed data. However, for the majority of the world's languages, and especially for low-resourced endangered languages, this approach is not a viable option because the amounts of texts that would be required – analyzed or not – are often not available. In many cases the language documentation work is the first source of any texts ever, although the increasing written use of many minority languages cannot be underestimated either. There have been successful projects which have built online corpora for a large of variety of these languages,[7] and it remains to be seen how they can be integrated to language documentation materials on these and related languages.

The older paradigm of language data analysis is the rule-based or grammar-based approach: the linguist writes a grammar rules in a specific format that is machine-readable, the formal grammar is then compiled into a program capable of analyzing (and eventually also generating) text input. There are several schools within the rule-based paradigm; the approach chosen by our projects is a combination of Finite-State Transducer (FST) technology for the morphological analysis, and Constraint Grammar (CG) for the syntactic analysis.

This approach has been tested with several written languages, for which it routinely provides highly robust analyses for unconstrained text input. We adapt the open preprocessing and analysis toolkit provided by Giellatekno (cf. Moshagen et al. 2013) for both written and spoken, transcribed language data. Since the chosen infrastructure is built for standard written languages, we have developed a set of conventions to convert our spoken language data into a "written-like" format, which is thus more easily portable into the Giellatekno infrastructure. First, we represent our spoken recordings in standardized orthography (with adaptations for dialectal and other sub-standard forms if needed). Second, we mark clause boundaries and use other punctuation marks as in written language, although surface text structuring in spoken texts is prosodic rather than syntactic and the alignment of our texts to the original record-

[6] http://giellatekno.uit.no, http://divvun.no

[7] http://web-corpora.net/wsgi3/minorlangs/

ing is utterance-based, rather than sentence-based. For specific spoken language phenomena, such as false starts, hesitations or self-corrections as well as for marking incomprehensible sections in our transcription, we use a simple (and orthography-compatible) markup adapted from annotation conventions commonly used in spoken language corpora.

Our transcribed spoken text data (using standard orthography) as well as any written text data are stored in the XML format provided by the multimedia language annotation program EUDICO Linguistic Annotator (ELAN, for short)[8] which allows audio and video recordings to be time-aligned with detailed, hierarchically organized tiers for transcriptions, translations and further annotations.

The annotation process contains the following steps:

1. **preprocessing**: a Perl script configured with a list of language-specific abbreviations that takes care of tokenization;
2. **morphosyntactic analysis**: an FST that models free and bound morpheme by means of linear (`lexc`) and non-linear rules (`twolc`) needed for word formation and inflection;
3. **disambiguation**: a formal grammar written in the CG framework.

The process of annotation enrichment in ELAN follows the usual analysis pipeline of the Giellatekno infrastructure. The string of each utterance is extracted from the `orthography`-tier, tokenized, then sent to the morphosyntactic analyzer, and finally to the disambiguation module. 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 and utterance. On the other hand, new XML element 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 four new tiers for `word`, `lemma`, `part-of-speech`, and `morphosyntactic description` that need to be generated and added to the initial structure, but 4xN, 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.

Since the ELAN files are in XML format, they can be both read and edited by humans with any text editor and accessed automatically by virtually any programming language. For the implementation of the script that links the ELAN data and the FST/CG, we decided to use Python because:

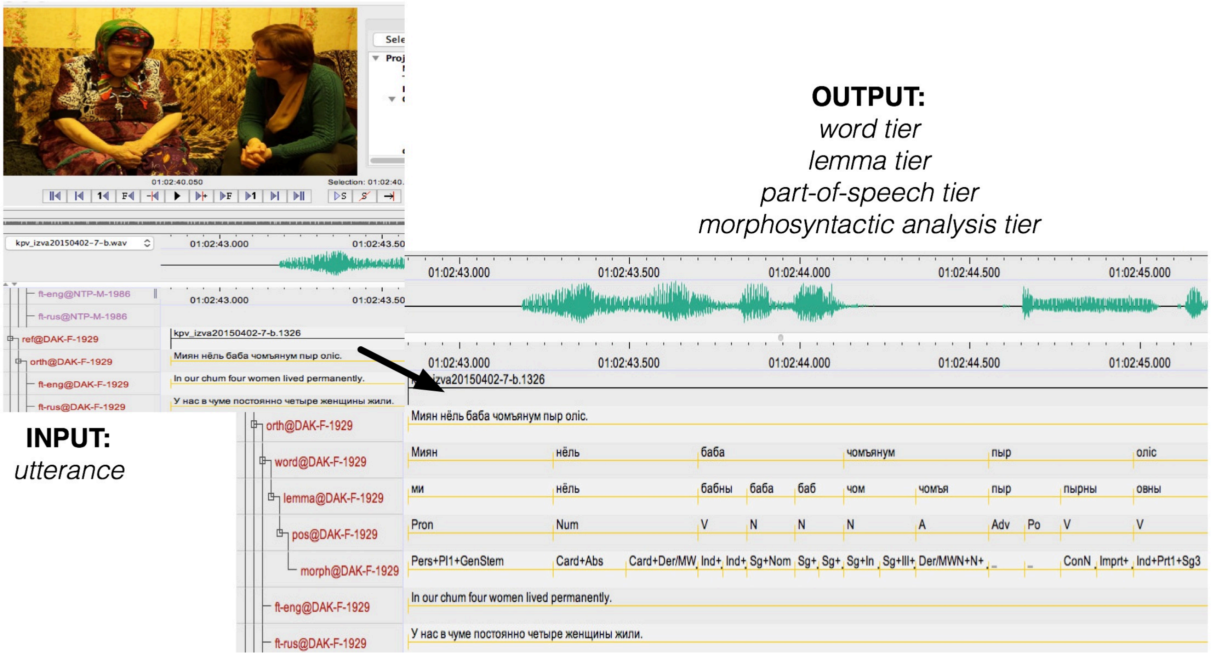
1. it is a flexible, interpreted programming language with support for multiple systems and platforms;
2. it is easy to read and to learn for even a novice programmer, which is perhaps the reason why it is often used for linguistic applications;
3. and finally, it offers XML processing support by means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

The input file for the whole process is an ELAN file lacking `word`, `lemma`, `part-of-speech`, and `morphological description` tiers. Thus, all tiers dependent on the `word`-tier are inserted dynamically (cf. Figure 1). For each speaker recorded in the ELAN file, the values of each utterance string from each individual `orth`-tier are extracted by the Python script and sent to the appropriate morphosyntactic analyzer.

After the FST has analyzed the word forms, has output the analyses, and the analyses are sent to disambiguation, the Python script parses the final output of the pipeline and restructures it when multiple analyses in a cohort are possible, that means when the disambiguation module could not disambiguate the output totally. A cohort is a word form along with all its possible analyses from the FST. 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.

Ambiguities in language analyses are quite common, but with FSTs in development for minority languages, they are even more frequent. Our

[8] `https://tla.mpi.nl/tools/tla-tools/elan`.



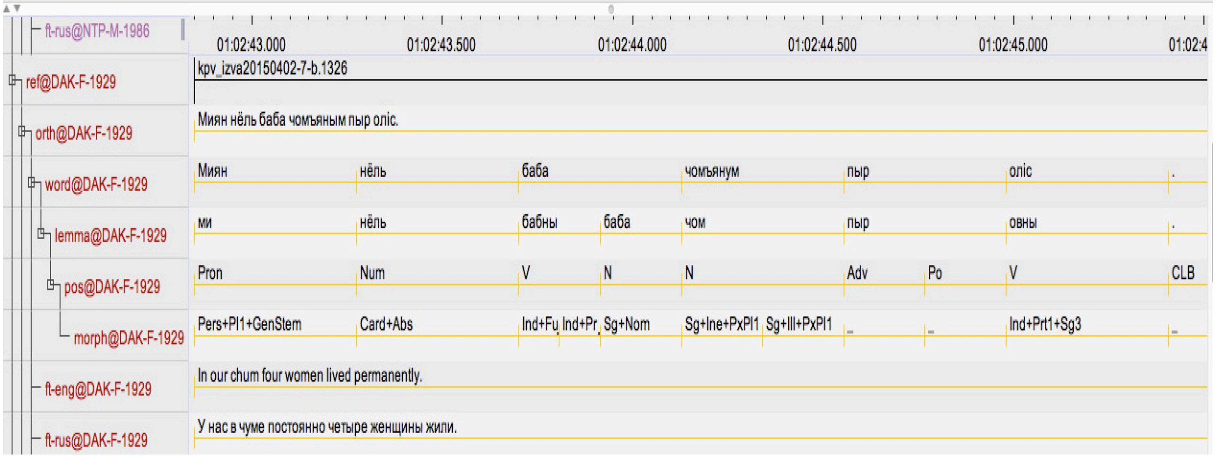
OUTPUT:
word tier
lemma tier
part-of-speech tier
morphosyntactic analysis tier

INPUT:
utterance

ft-eng@NTP-M-1986
ft-rus@NTP-M-1986
ref@DAK-F-1929 kpv_izva20150402-7-b.1326
orth@DAK-F-1929 Миян нёль баба чомъянум пыр олiс.
ft-eng@DAK-F-1929 In our chum four women lived permanently.
ft-rus@DAK-F-1929 У нас в чуме постоянно четыре женщины жили.

orth@DAK-F-1929 Миян нёль баба чомъянум пыр олiс.
word@DAK-F-1929 Миян | нёль | баба | чомъянум | пыр | олiс
lemma@DAK-F-1929 ми | нёль | бабны | баба | баб | чом | чомъя | пыр | пырны | овны
pos@DAK-F-1929 Pron | Num | V | N | N | N | A | Adv | Po | V | V
morph@DAK-F-1929 Pers+Pl1+GenStem | Card+Abs | Card+Der/MW | Ind+ | Ind+ | Sg+Nom | Sg+ | Sg+ | Sg+In | Sg+Ill+ | Der/MWN+N+ | _ | _ | ConN | Imprt+ | Ind+Prt1+Sg3
ft-eng@DAK-F-1929 In our chum four women lived permanently.
ft-rus@DAK-F-1929 У нас в чуме постоянно четыре женщины жили.

Figure 1: The result of the first two processing steps – tokenization and morphosyntactic analysis for the Komi sentence *Миян нёль баба чомъянум пыр олiс.* ‘In our chum (tent) four women lived permanently.’



ft-rus@NTP-M-1986
ref@DAK-F-1929 kpv_izva20150402-7-b.1326
orth@DAK-F-1929 Миян нёль баба чомъяным пыр олiс.
word@DAK-F-1929 Миян | нёль | баба | чомъянум | пыр | олiс | .
lemma@DAK-F-1929 ми | нёль | бабны | баба | чом | пыр | овны | .
pos@DAK-F-1929 Pron | Num | V | N | N | Adv | Po | V | CLB
morph@DAK-F-1929 Pers+Pl1+GenStem | Card+Abs | Ind+Fu | Ind+Pr | Sg+Nom | Sg+Ine+PxPl1 | Sg+Ill+PxPl1 | _ | _ | Ind+Prt1+Sg3 | _
ft-eng@DAK-F-1929 In our chum four women lived permanently.
ft-rus@DAK-F-1929 У нас в чуме постоянно четыре женщины жили.

Figure 2: The result of the last processing step – (partial) disambiguation of morphosyntactic ambiguous analyses for the same Komi sentence as in Figure 1: for instance, for the word form *чомъянум* the contextually appropriate `noun` reading with lemma *чом* is chosen while the `adjective` reading with lemma *чомъя* is discarded.

plan is to further develop the extant disambiguation module in a similar way as for Northern Saami (cf. Antonsen, Huhmarniemi, et al. 2009).

Note that the lack of a higher-level analysis often leads to cases of ambiguity concerning the morphological analysis, i.e., multiple analyses for one and the same word form. As already mentioned, for the disambiguation of these homonyms, we use CG, which takes the morphologically analyzed text as its input, and ideally only returns the appropriate reading, based on the immediate context of a specific word form–analysis pair. CG is a language-independent formalism for morphological disambiguation and syntactic analysis of text corpora developed by Karlsson (1990) and Karlsson et al. (1995). Since we use the standard analysis pipeline offered by the Giellatekno's infrastructure, we use the CG implementation vislcg3,[9] which is freely available.

The CG analysis can be enriched with syntactic functions and dependency relations if all underlying grammatical rules are described sufficiently. Since the output of a CG analysis is a dependency structure for a particular sentence, the output may also be converted into phrase structure representations.

Our work with the CG description of Komi is only at the beginning stage. To be completed, it would likely need to include several thousand rules. However, the experience of other Giellatekno projects working with CG shows that some months of concentrated work can result in a CG description that can already be implemented in a preliminary tagger useful for lexicographic work as well as for several other purposes. For instance, the rather shallow grammar parser for Southern Saami described by Antonsen and Trosterud (2011) includes only somewhat more than 100 CG rules, but already results in reasonably good lemmatization accuracy for open class parts-of-speech. This means that the approach is readily adjustable to language documentation projects with limited resources. Furthermore, CG rules can potentially be ported from one language to another, e.g. the rule for disambiguating the connegative verb in Komi would also work in several other Uralic languages.

Komi is an agglutinative language, so with morphologically more complex forms the homonymy is rare. However, there are specific forms which tend to collide with one another, especially with some verb and noun stems which happen to be identical. The monosyllabic stems have canonical shape CV(C), and this simple structure is one reason that makes morphologically unmarked forms to be homonymous: nominative singular nouns, and some of the imperative and connegative forms in the verbal paradigm.

4 Post-Processing of Corpus Data in ELAN

Note 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 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.

However, since our approach aims at separating the annotation tool (i.e. the FST/CG-based tagger) from the non-annotated corpus data (i.e. the transcription), we are relying on the ELAN GUI only until the transcription (and translation) of the original recordings is done in this program. The ELAN-FST/CG interaction does not depend on the ELAN tool, but only on the data in ELAN XML. Once the corpus is extended by new transcripts (or existing transcripts are corrected) or the FST/CG rules are refined, the whole corpus will be analyzed anew. This will simply overwrite the previous data in the relevant ELAN tiers.

In this way, we will release new annotated corpus in regular intervals and make them available at TLA in Nijmegen, where we archive and make available our project data. One significant advantage of working with ELAN XML is that it can be accessed through the online tools Annotation Explorer (ANNEX, for short)[10] and TROVA[11], which are basically online GUIs of the same search engines built into the ELAN tool. 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

[9] https://visl.sdu.dk/cg.html

[10] `https://tla.mpi.nl/tools/tla-tools/annex`

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

same was as within ELAN itself. As a practical benefit, the integration of TROVA and the whole infrastructure at TLA makes it easy to control corpus access, something that regularly 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.

5 Summary

In this paper, we describe a project at the interface between language technology and language documentation. Our aim is to overcome manual annotation of our language documentation corpora. We achieve this by implementing Finite State Transducers and Constraint Grammar for rule-based morphosyntactic tagging and disambiguation. The different modules in our annotation tool interact directly with our corpus data, which is consistently structured in ELAN XML.

Since Constraint Grammar can be simply chained, we plan not only to extend and improve the current disambiguation module but also to implement Constraint Grammar for further, syntactic analysis, to achieve fully disambiguated dependency treebanks. The relevant work will be carried out as part of the project "Language Documentation meets Language Technology: The Next Step in the Description of Komi" which already started in 2017.

While the rule-based morphosyntactic modeling is not state of the art in contemporary NLP, it does have significant advantages specifically in *endangered* language documentation:

1. the results of the automatic tagging are *exceptionally precise* and cannot normally be reached with statistical methods applied on the relatively small corpora endangered language documentation normally creates;

2. while incrementally formulating rules and testing them on the corpus data, we are not only creating a tool but producing a *full-fledged grammatical description based on broad empirical evidence* at the same time;

3. and last but not least, our work will eventually also help develop new (I)CALL technology, i.e. *(intelligent) computer-assisted language learning* systems for the languages we work on.

One relatively simple example for the latter is the direct implementation of our formalized morphosyntactic description in Voikko spell-checkers, which is a easily done with the Giellatekno/Divvun infrastructure setup and which is an highly important tool to support future writers of the endangered language.

Last but not least, since the official support and language planning activities are significant at present for Komi and some of the other languages we are working on, these languages are increasingly used in spoken and written form. Better adaptation of computational technology by language documenters will eventually be necessary in order to annotate and make efficient use of the ever increasing amount of available data.

Ultimately, our work will therefore also contribute to future language planning, language development, and language vitalization.

Acknowledgments

This paper describes on-going work by the Izhva Komi Documentation Project, which is funded by the Kone Foundation between 2014–2016 and 2017–2020 and embedded in a larger research group[12] working with the documentation and description of several endangered languages of the Barents Sea Area in Northeastern Europe. The present article builds on and continues our preliminary research published in Blokland, Gerstenberger, et al. 2015; Gerstenberger et al. 2017a; Gerstenberger et al. 2017b. We would like to thank the other co-authors of these papers, Rogier Blokland, Marina Fedina, and Joshua Wilbur for their ongoing collaboration. We also want to express our gratitude to Trond Trosterud and Jack Rueter for continuous valuable feedback. Last but not least, important ideas for our research were developed while writing the application for the long-term project INEL[13] by Beáta Wagner-Nagy together with Michael Rießler and The Hamburg Center for Language Corpora (University of Hamburg).

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

[13] https://inel.corpora.uni-hamburg.de

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.

Antonsen, Lene and Trond Trosterud (2011). “Next to nothing. A cheap South Saami disambiguator”. In: *Proceedings of the NODALIDA 2011 Workshop Constraint Grammar Applications, May 11, 2011 Riga, Latvia*. Ed. by Eckhard Bick, Kristin Hagen, Kaili Müürisep, and Trond Trosterud. NEALT Proceedings Series 14. Tartu: Tartu University Library, pp. 1–7. url: `http://hdl.handle.net/10062/19296`.

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 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`.

Cox, Christopher (2011). “Corpus linguistics and language documentation. Challenges for collaboration”. In: *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*. Ed. by John Newman, Harald Baayen, and Sally Rice. Amsterdam: Rodopi, pp. 239–264.

Fedina, Marina, Enye Lav, Dmitri Levchenko, Ekaterina Koval, and Inna Nekhorosheva (2016–2017). *Nacional’nyj korpus komi jazyka*. Syktyvkar: FU-Lab. url: `http://komicorpora.ru`.

Gerstenberger, Ciprian, Niko Partanen, Michael Rießler, and Joshua Wilbur (2017a). “Instant annotations. Applying NLP methods to the annotation of spoken language documentation corpora”. In: *Proceedings of the 3rd International Workshop on Computational Linguistics for Uralic languages. Proceedings of the workshop*. Ed. by Tommi A. Pirinen, Michael Rießler, Trond Trosterud, and Francis M. Tyers. ACL anthology. Baltimore, Maryland, USA: Association for Computational Linguistics (ACL). url: `http://aclweb.org/anthology/`. In press.

– (2017b). “Utilizing language technology in the documentation of endangered Uralic languages”. In: *Northern European Journal of Language Technology*: *Special Issue on Uralic Language Technology*. Ed. by Tommi A. Pirinen, Trond Trosterud, and Francis M. Tyers. url: `http://www.nejlt.ep.liu.se/`. In press.

Hausenberg, Anu-Reet (1998). “Komi”. In: ed. by Daniel Abondolo. Routledge Language Family Descriptions. London: Routledge, pp. 305–326.

Himmelmann, Nikolaus (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`.

Karlsson, Fred (1990). “Constraint Grammar as a framework for parsing unrestricted text”. In: *Proceedings of the 13th International Conference of Computational Linguistics*. Ed. by Hans Karlgren. Vol. 3. Helsinki, pp. 168–173.

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, and Arto Anttila, eds. (1995). *Constraint Grammar. A language-independent system for parsing unrestricted text*. Natural Language Processing 4. Berlin: Mouton de Gruyter.

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`.

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`.

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`.

Tiedemann, Jörg, Johanna Nichols, and Ronald Sprouse (2016). "Tagging Ingush. Language technology for low-resource languages using resources from linguistic field work". In: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH). Osaka, Japan, December 11–17 2016*, pp. 148–155. url: `https://www.clarin-d.de/joomla/images/lt4dh/pdf/LT4DH20.pdf`.

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.

Waldenfels, Ruprecht von, Michael Daniel, and Nina Dobrushina (2014). "Why standard orthography? Building the Ustya River Basin Corpus, an online corpus of a Russian dialect". In: *Computational linguistics and intellectual technologies. Proceedings of the Annual International Conference «Dialogue» (2014)*. Moskva: Izdatel'stvo RGGU, [720–729]. url: `http://www.dialog-21.ru/digests/dialog2014/materials/pdf/WaldenfelsR.pdf`.