Instant Annotations – Applying NLP Methods to the Annotation of Spoken Language Documentation Corpora

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

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

Michael Rießler†
ENS & PSL Research University / University of Freiburg
LaTTiCe / 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

*The order of the names of the authors is alphabetical.

†Michael Rießler's contribution to this work has received support of TransferS (laboratoire d'excellence, program "Investissements d'avenir" ANR-10-IDEX-0001-02 PSL* and ANR-10-LABX-0099).

The 3rd International Workshop for Computational Linguistics of Uralic Languages, pages 25–36,
St. Petersburg, Russia, 23–24 January 2017.

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 in Freiburg and in collaboration with Hamburg, Syktyvkar, Tromsø and Uppsala. Our projects work in the endangered language documentation framework and 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 NLP methods in language documentation – specifically rule-based morphological and syntactic analyzers – helps us to create more systematically annotated corpora, rather than eclectic data collections. We propose a step-by-step approach to reach higher-level annotations by using and improving truly computational methods. Ultimately, the spoken corpora created by our projects will be useful for scientifically significant quantitative investigations on these languages in the future.

1 Introduction

Endangered language documentation (aka documentary linguistics) aims at the provision of long-lasting, comprehensive, multifaceted and multipurpose records of linguistic practices characteristic of a given speech community [1, 2, 3, 4]. The field 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 significant corpora. This is despite the fact that the relevant computational methods and tools are well-known from corpus-driven linguistic research on larger written languages and are even applied to spoken varieties of these languages.

Although relatively small endangered languages are increasingly in the focus of computational linguistic research (see especially Giellatekno for Northern Saami [5] and other languages, a different approach is [6]), 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 for these minority languages and include higher-level morphosyntactic annotations. However, they represent a limited range of text genres, typically including formal styles, and they include large portion of translations from the relevant majority languages.[1]

[1]The metadata provided with the Northern Saami written corpus at Giellatekno [7] suggests that the

On the other hand, researchers working in the framework of endangered language documentation (so-called "documentary linguistics"), 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 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),[2] FieldWorks Language Explorer (or FLEx, for short)[3] 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. Two examples of comparably large corpora created in this approach and supposedly even exceeding the number of 150,000 tokens, are the Nganasan corpus described by [9][4] and the corpus of Forest and Tundra Enets [10].[5] Typically, such spoken corpora are smaller, as is the case for the annotated corpus of spoken Beserman Udmurt comprising 65,000 tokens[6], the annotated corpora of spoken Eastern Khanty and Southern Selkup [11],[7] and the annotated corpora of Tundra Nenets and Northern Khanty.[8] 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 transcription. Examples for recent projects which use phonemic transcription instead of an orthographic standard are the Khanty, Tundra Nenets, and Udmurt documentations described by [12] (in the current proceedings), the Northern Selkup documentation currently carried out as part of the INEL project,[9] as well as the corpora mentioned in the preceding paragraph. Note that most Uralic languages (or at

portion of non-translated texts is rather high, which would be against our earlier statement in [8].

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

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

[4] Inferred from [9], who do not quantify the corpus in terms of tokens but mention the inclusion of "59 texts", and the description of somehow related data at `http://www.iling-ran.ru/gusev/Nganasan/`, mentioning "14,928 sentences (including approximately 28,000 types)" [our translation].

[5] Olesya Khanina, p.c.

[6] `http://urn.fi/urn:nbn:fi:lb-2015081401`

[7] An indication of the actual size, in terms of texts, sentences tokens or the like, is not given.

[8] `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.

[9] `https://inel.corpora.uni-hamburg.de/?page_id=173`; according to the INEL project application, which was co-authored by one of the current authors; a general introduction to INEL is [13].

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] 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 materials like these in automatic corpus annotation has already been reported as a well working approach [15].

Particularly when basic phonological and morphological descriptions are already available 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 [8]).

We suggest the following two main principles, which we have begun implementing consistently in our own documentation projects on languages from the Permic and Saamic branches: (1) Use an orthography-based transcription system; this not only allows quicker and more efficient transcription of field recordings, but it makes it possible to easily integrate all available (digitized) printed texts into the corpus. 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. (2) Apply computer-based methods as much as possible in creating higher-level annotations of the compiled corpus data.

The examples in our paper are taken specifically from Komi-Zyrian, an endangered Uralic language. Other endangered Uralic languages we work on at present are Akkala Saami, Kildin Saami, Pite Saami, Skolt Saami and Ter Saami. We present our work-in-progress concerning the application of rule-based morphological tagging and syntactic disambiguation in order to automatically create higher-level corpus annotations. In this, our aim is to challenge and further develop current approaches at the interface between computational, descriptive and documentary linguistics of endangered languages.

[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 Project [14].

2 Spoken corpus annotation

The dominating paradigm within computational linguistics is based on statistical methods and training a computer to understand the behavior of natural language by means of presenting it 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 typically not available. The competing paradigm is a rule-based ("grammar-based") analysis: a linguist writes a machine-readable version of the grammar, and compiles it 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 uses a combination of finite-state transducer technology for morphological analyses, and Constraint Grammar for the syntactic analyses.

This approach has been tested for several written languages, and it routinely provides highly robust analyses for unconstrained text input. We adapt the open-source preprocessing and analysis toolkit provided by the Giellatekno project [16][12] for both written and spoken, transcribed language data. Since the Giellatekno infrastructure is built for standard written languages, we have developed a set of conventions for converting our spoken language data into a "written-like" format that 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 when needed), rather than in phonemic transcription (this is unlike many other endangered language documentation projects). Second, we mark clause boundaries and use other punctuation marks as in written language, even though surface text structuring in spoken texts is prosodic rather than syntactic and the alignment of our texts to the original recording is utterance-based, rather than sentence-based. For specific spoken-language phenomena, such as false starts, hesitations or self-corrections as well as when marking incomprehensible sections in our transcription, we use a simple (and orthography-compatible) markup adapted from annotation conventions commonly used in spoken language corpora.[13] Different resources on endangered languages have typically used different transcription conventions and orthographies, and essentially our approach using orthography is based on the idea that we should select a single system for transcriptions. The current orthography is the most established one of the different variants, and is used for the the largest amount of available texts. The orthographies on the languages we work with are relatively phonemic, al-

[12]Giellatekno, The Center for Saami Language Technology (University of Tromsø), `http://giellatekno.uit.no/`

[13]Our convention is based on HIAT [17], but is much simpler and only includes a few rules.

though the Cyrillic writing system and borrowed Russian conventions lead to a few additional cosmetic details. However, it still represents the underlying phoneme level very well, and any texts using more narrow transcriptions can always be converted to the orthography; at the same time, virtually all other transcription systems used can be transliterated into the orthographic representation as they generally still adhere to the same phoneme level. In addition, using orthography makes our transcriptions readily accessible to the language community because speakers are used to reading in orthography; this even makes it easier to employ native speakers to work on transcribing the segmented audio data.

The annotation process works on three levels:

(1) The first level is a preprocessor which tokenizes the orthographic transcription.

(2) The second level is a morphological analyzer, programmed as a finite-state transducer (FST) for modeling free and bound morphemes as well as linear and nonlinear rules according to which morphemes combine in word formation and inflection: 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 transducer is fed to a so-called Two-Level-Morphology (TWOL) component [18] used for handling complex suprasegmental morphophonological rules (which are particularly characteristic of the Saamic languages, but much less so of Komi).

(3) The third level is a syntactic analyzer-disambiguator, written as a set of rules following Constraint Grammar (CG). 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, which is of course problematic since any given token has a single correct morphological analysis. For the syntactic disambiguation of these homonyms, we use CG, which takes the morphologically analyzed text as its input, and ideally only returns the appropriate reading. CG is a language-independent formalism for morphological disambiguation and syntactic analysis of text corpora developed by [19, 20]. 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, cf. the example in Figure 1. Similar to other projects using the Giellatekno toolkit, we use VISLCG-3 for the compilation of the manually written CG rules [21].[14]

The following examples illustrate a possible case of homonymy to be disambiguated after the FST morphological analysis.

1. сёй : сёй+N+Sg+Nom

[14]VISLCG-3 is an improved version of VISL, documented at `http://beta.visl.sdu.dk/cg3.html`.

2. сёй : сёйны+V+ConNeg

3. сёй : сёйны+V+Imprt+Sg2

Here, the token to be analyzed is сёй. The analyzer spells out the possible lemmas сёй 'clay' and сёйны 'to eat' followed by the possible part-of-speech and morphological category tags for a total of three theoretically possible readings. One example of a (relatively simple) syntactic rule used in the disambiguation of a token сёй would be:

- IFF: ConNeg if Neg to the left

This rule would apply when the token сёй follows a negation verb inside running text, thus selecting the second analysis as the correct one in such a case.

Our work with the CG description of Komi is still at an initial stage. For the Saamic languages, we have not started work with CG yet. 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 [22] 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 adaptable for language documentation projects with limited resources. Furthermore, CG rules can potentially be ported from one language to another, e.g. the rule described above for disambiguating the connegative verb in Komi would also work for several other Uralic languages.

3 Summary

Although endangered language documentation has a focus on multi-modal speech corpora and uses data from small orally transmitted languages, the relevant research is in essence similar to corpus building of any other non-endangered and/or written language. However, endangered language documentation does not seem to be well informed by common theories known from "non-endangered corpus linguistics" and typically does not even consider using computational methods for corpus annotation and the creation of qualitatively and quantitatively more significant corpora. Why do the majority of endangered language documentation projects still rely entirely on non-automated methods if NLP has already been applied successfully to very small

languages?[15] A possible answer is that many linguists working with language documentation come from comparative and descriptive linguistics and prefer qualitative methods. The approach described in our paper tries to consistently apply proven methods from NLP in endangered language documentation and potentially even in endangered language description.

While rule-based morphosyntactic modeling is initially time-consuming (at the development stage), it does have significant advantages: (1) the results of automatic tagging are exceptionally precise and consistent, and – obviously – automatic; (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; and last but not least (3) our work can eventually even help develop new language technology for computer-aided teaching and writing. For instance, our FST descriptions are implemented in the creation of spell-checkers using the Giellatekno toolkit.

Due to the fact that significant official support and language planning activities currently exist for Komi as well as 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 researchers working in the field of language documentation will in the long run become necessary in order to more efficiently annotate and make effective use of the increasing amount of data available.

Whereas the rule-based methods described in this paper have already been successfully used with *written* varieties of Komi and Saamic languages, our paper describes their application specifically to *spoken* varieties. This approach is a novelty in the field of language documentation and computational linguistics for small Uralic languages and not at all a trivial task. It requires innovative research for several reasons: (1) specific spoken-language phenomena (false starts, self-corrections, incomprehensible speech, etc.) marked in transcriptions need to be pre-processed systematically; (2) additional morphological and syntactic rules need to be introduced to process linguistic variation characteristic of spoken varieties; and last but not least (3) our corpus data is often not monolingual, but instead includes a significant amount of borrowings from Russian or other relevant majority languages for other projects as well as code-switching into these languages.

This last point is worth explaining in more detail because it addresses a potential (and obviously necessary) direction to take in the future of automatic corpus annotation of spoken Uralic language data. As mentioned above, we use orthographic transcriptions consistently, even for non-target languages present in our corpus data,

[15]This question was asked by Arienne Dwyer in a recent project description, `http://www.nsf.gov/awardsearch/showAward?AWD_ID=1519164`.

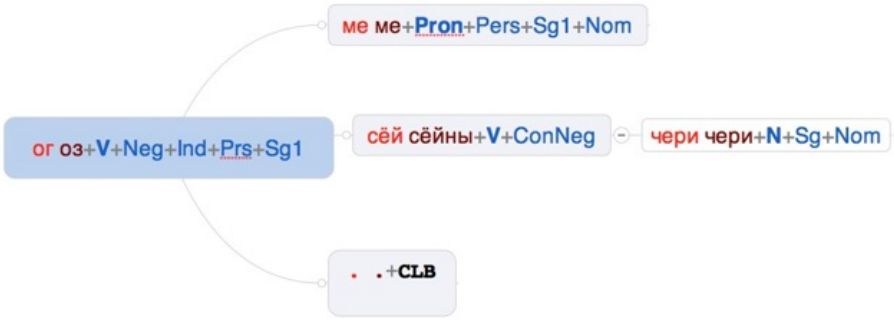

Figure 1: Dependency tree for the Komi sentence *Ме чери ог сёй.* "I don't eat fish" after disambiguation

i.e., switches to Russian and Tundra Nenets. Since analyzers for Russian and Tundra Nenets are also available in the Giellatekno infrastructure, we can run multiple language analysis easily. This is a direct benefit and consequence of adopting the pre-existing Giellatekno infrastructure and of using orthographies in transcription. As a result, it becomes possible to automatically detect the parts of our corpus where multiple languages occur. Because of the rule-based approach requires all combinations of free and bound morphemes to be detected in the corresponding lexica and rules, this works best when switches between languages are indisputable (rather than ad hoc borrowings or other hybrid forms). While the handling of mixed language data in our corpora is not yet entirely worked out, we can already use this approach for concrete tasks, such as for assigning language tags to different recordings. Improved methods for automatically detecting code-mixing and code-switching and then merging the resulting analyses, are in the works.

References

[1] Nikolaus Himmelmann. Language documentation. What is it and what is it good for? In Jost Gippert, Ulrike Mosel, and Nikolaus Himmelmann, editors, *Essentials of Language Documentation*, number 178 in Trends in Linguistics. Studies and Monographs, pages 1–30. Mouton de Gruyter, Berlin, 2006.

[2] Anthony C. Woodbury. Language documentation. In Peter K. Austin and Julia Sallabank, editors, *The Cambridge handbook of endangered languages*, Cam-

bridge handbooks in language and linguistics, pages 159–186. Cambridge University Press, Cambridge, 2011.

[3] 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.

[4] Peter K. Austin. Language documentation in the 21st century. *JournaLIPP*, 3:57–71, 2014. URL: http://lipp.ub.lmu.de/article/download/190/83.

[5] Trond Trosterud. Grammar-based language technology for the Sámi languages. In *Lesser used Languages & Computer Linguistics*, pages 133–148. Europäische Akademie, Bozen, 2006.

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

[7] *SIKOR. UiT The Arctic University of Norway and the Norwegian Saami Parliament's Saami text collection, Version 08.12.2016*. Tromsø, 2016. URL: http://gtweb.uit.no/korp.

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

[9] Wagner-Nagy Beáta and Sándor Szeverényi. Linguistically annotated spoken Nganasan corpus. *Tomsk Journal of Linguistics and Anthropology*, 2:25–33, 2015.

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

[11] Andrey Filchenko and Balthasar Bickel. Comprehensive documentation and analysis of two endangered Siberian languages. Eastern Khanty and Southern Selkup. In *The Endangered Language Archive (ELAR)*. SOAS University of London, London, [n.d.]–2017. URL: https://elar.soas.ac.uk/Collection/MPI43298.

[12] Eszter Simon and Nikolett Mus. Languages under the influence. Building a database of Uralic languages. In Tommi A. Pirinen, Francis M. Tyers, and Trond Trosterud, editors, *Third International Workshop on Computational Linguistics for Uralic languages, 23rd January, Saint Petersburg/Russia. Proceedings of the workshop*. 2017. In press.

[13] Beata Wagner-Nagy, Hanna Hedeland, Timm Lehmberg, and Michael Rießler. INEL. Eine Infrastruktur zur Dokumentation indigener nordeurasischer Sprachen. In *Konferenz "Forschungsdaten in den Geisteswissenschaften (FORGE 2015)". 15. bis 18. September 2015 an der Universität Hamburg*. Lecture2Go. Projekt Geisteswissenschaftliche Infrastruktur für Nachhaltigkeit (gwin), Hamburg, 2015. URL: https://lecture2go.uni-hamburg.de/l2go/-/get/v/18306.

[14] Heidi Jauhiainen, Tommi Jauhiainen, and Krister Lindén. The Finno-Ugric Languages and The Internet Project. In Tommi A. Pirinen, Francis M. Tyers, and Trond Trosterud, editors, *First International Workshop on Computational Linguistics for Uralic Languages, 16th January, 2015, Tromsø, Norway. Proceedings of the workshop*. Volume 2015, number 2 in Septentrio Conference Series, pages 87–98. The University Library of Tromsø, Tromsø, 2015. DOI: 10.7557/5.3471.

[15] Timofey Arkhangelskiy and Maria Medvedeva. Developing morphologically annotated corpora for minority languages of russia. In Sandra Kübler and Markus Dickinson, editors, *Proceedings of Corpus Linguistics Fest 2016. Bloomington, IN, USA, June 6-10, 2016*. Pages 1–6, 2016.

[16] 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*, number 16 in NEALT Proceedings Series, pages 343–352, 2013.

[17] Jochen Rehbein, Thomas Schmidt, Bernd Meyer, Franziska Watzke, and Annette Herkenrath. *Handbuch für das computergestützte Transkribieren nach HIAT*, number 56 in Arbeiten zur Mehrsprachigkeit, Folge B. Universität Hamburg, Hamburg, 2004. URL: http://www.exmaralda.org/files/azm_56.pdf.

[18] 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.

[19] Fred Karlsson. Constraint Grammar as a framework for parsing unrestricted text. In Hans Karlgren, editor, *Proceedings of the 13th International Conference of Computational Linguistics*. Volume 3, pages 168–173. Helsinki, 1990.

[20] Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Arto Anttila, editors. *Constraint Grammar. A language-independent system for parsing unrestricted text*, number 4 in Natural Language Processing. Mouton de Gruyter, Berlin, 1995.

[21] Tino Didriksen. *Constraint grammar manual. 3rd version of the CG formalism variant*. GrammarSoft ApS, 2007–2016. URL: `http://visl.sdu.dk/cg3/vislcg3.pdf`.

[22] Lene Antonsen and Trond Trosterud. Next to nothing – a cheap South Saami disambiguator. In Eckhard Bick, Kristin Hagen, Kaili Müürisep, and Trond Trosterud, editors, *Proceedings of the NODALIDA 2011 Workshop Constraint Grammar Applications, May 11, 2011 Riga, Latvia*, number 14 in NEALT Proceedings Series, pages 1–7. Tartu University Library, Tartu, 2011. URL: `http://hdl.handle.net/10062/19296`.

[23] Tommi A. Pirinen, Francis M. Tyers, and Trond Trosterud, editors. First International Workshop on Computational Linguistics for Uralic Languages, 16th January, 2015, Tromsø, Norway. Volume 2015. (2) in Septentrio Conference Series. The University Library of Tromsø, Tromsø, 2015.