A pseudonymisation method for language documentation corpora: An experiment with spoken Komi

Niko Partanen
University of Helsinki
Helsinki, Finland
niko.partanen@helsinki.fi

Rogier Blokland
Uppsala University
Uppsala, Sweden
rogier.blokland@moderna.uu.se

Michael Rießler
University of Eastern Finland
Joensuu, Finland
michael.riessler@uef.fi

Abstract

This article introduces a novel and creative application of the Constraint Grammar formalism, by presenting an automated method for pseudonymising a Zyrian Komi spoken language corpus in an effective, reliable and scalable manner. The method is intended to be used to minimize various kinds of personal information found in the corpus in order to make spoken language data available while preventing the spread of sensitive personal data about the recorded informants or other persons mentioned in the texts. In our implementation, a Constraint Grammar based pseudonymisation tool is used as an automatically applied shallow layer that derives from the original corpus data a version which can be shared for open research use.

Teesiq

Seo artikli tutvustas vahtsõt ja loovat piirdmiisi grammatiga (PG) formalismõ pruuk´mist. Taas om metod´, kon PG pruugitas tuusjaos, et süräkomi kõnõldu keele korpusõ lindistuisi saassiq tegüsähe, kimmähe ja kontrol´misõvõimalusõga vaṕonimmiga käkkiq. Seo metod´ om tett, et korpusõn saassiq kõnõlõjidõ andmit nii pall´o vähembäs võttaq, ku või, ja et tulõmit saassiq kergehe käsilde kontrolliq. Mi plaani perrä pruugitas taad ku automaatsõt kihti, miä tege korpusõ säändses, et taad või kergehe uuŕmisõ jaos jakaq.

1 Introduction

The research presented in this paper is predominantly relevant for documentary linguistics, aiming at the creation of a "lasting multipurpose record of a language" (Himmelmann, 2006, 1), while applying a computational linguistic approach to an endangered Uralic language. Specifically, we are developing an automated method for pseudonymising the textual representation of a spoken language corpus in order to make the corpus data publishable while 1) preventing the spread of sensitive personal data, 2) overcome manual work in the process to the extent possible, and 3) keeping the pseudonymised data as one – more openly distributed – version of the original – and less openly distributed – data, rather than destroying the latter by overwriting or cutting away parts of them.

To our knowledge, this is a novel approach in documentary linguistics, which so far seems to rely mostly on manual methods for pseudonymising (or anonymising) corpus data or bypasses the problem, typically by generally applying very restrict access protocols to corpus data preventing them from being openly published.

Computational linguistic projects aiming at corpus building for endangered languages, on the other hand, are rarely faced the problem of personal data protection because their corpora typically originate from written texts, which either are openly available to begin with or for which access rights have been cleared before the work with corpus building starts. Different from written-language data, corpora resulting from fieldwork-based spoken-language documentation invariably contain large amounts and various kinds of personal data. The reason for this is that documentary linguists transcribe authentic speech samples meant to be re-used in multidisciplinary research beyond structural linguistics.

Proceedings of the 6th International Workshop on Computational Linguistics of Uralic Languages, pages 1–8
Wien, Austria, January 10 — 11, 2020.
https://doi.org/10.18653/v1/FIXME

Typically, recordings are done with members of small communities, where most individuals know each other, and common topics include oral histories about places, persons or events inevitably including personal information about the informants itself or other individuals.

Although the recorded speaker's informed consent to further processing and re-using the speech sample needs to be at hand in any case, fieldwork-based recordings nearly always include information which should not be made entirely openly available. Simultaneously, it is in the natural interest of documentary linguistics to make as many materials as widely available as possible. The approach discussed in this paper consequently attempts to find a suitable middle way in presenting our Zyrian Komi corpus to a wider audience, in this case mainly researchers such as linguists and anthropologists, whilst ensuring that the privacy of individual speakers is respected and the risk of miss-using personal data can be excluded.

Best practice recommendations related to the problem of personal data protection in Open Science are currently evolving (cf. Seyfeddinipur et al., 2019), although relevant issues have been under discussion for a while already in the context of in Documentary Linguistics. The conventions and technical solutions described on this study have been developed in a specific situation, where the goal has been to publish and archive the corpus in the Language Bank of Finland (Blokland et al., 2020), and can hopefully contribute to solving at least some of the many challenges still remaining open.

Since GDPR-related[1] research practices are still evolving in Finland and there are as yet no clear guidelines, it is currently problematic to make audiovisual research materials containing identifiable personal information available. As there are fewer limitations when the material is anonymised or pseudonymised, we have explored this as a solution: a version of a corpus that does not contain identifiable personal information can be openly shared much more easily. The current approach we are considering is to share the current corpus with academic users through the Korp interface (Ahlberg et al., 2013) under so-called ACA conditions. This ensures that the users are authenticated as members of the academic community, and their identity is known. At this level, however, they can only access the versions that have passed through the pseudonymisation system described in this paper. We refer to this method as pseudonymisation, since the actual identity information is not discarded permanently from existence. We just derive a version where the personal data is minimised, and provide that to one particular user group.

We think this approach can be a satisfactory middle way to make the data accessible to the scientific community without needlessly revealing the personal information of individual speakers. We aim, however, to make the complete corpus available for research use with a specific application procedure, as is common with language documentation corpora in other archives.

These methods to share the corpora primarily serve academic users in Europe, but not the community itself, and to resolve this issue our colleagues from the community have also made a selection of the corpus available as a 'community edition' at a website videocorpora.ru[2], which is maintained and curated by FU-Lab (the Finno-Ugric Laboratory for the Support of Electronic Representation of Regional Languages in Syktyvkar, Russia). This, however, though trilingual (Zyrian Komi, Russian, English), is designed mainly from the point of view of and for community members, containing edited video versions aimed at an uncomplicated user experience, and does not (and is not primarily supposed to) satisfy the needs of many academic users.

In order to reliably pseudonymise the transcriptions in a version of the corpus aimed at research use we have developed a workflow that uses existing rule-based NLP for Zyrian Komi. The approach consists of performing an analysis on all running text, with various strategies to manipulate and filter out proper nouns, such as person names and toponyms. This allows us to keep some of the naturalness of the running text, while removing and changing easily identifiable content. Another benefit of our approach is that it lets us show which recordings contain which types of personal information.

Our method uses Finite-State Morphology (henceforth FST), specifically HFST (Lindén et al., 2009), and Constraint Grammar (henceforth CG) (Karlsson, 1990; Karlsson et al., 2011), specifically the CG-3 version (Bick and Didriksen, 2015), and is applied to the corpus using a uralicNLP Python package (Hämäläinen, 2019). The use of rule-based NLP methods is, in our opinion, highly desirable in this context as we can thus

[1]The General Data Protection Regulation (EU) 2016/679

[2]http://videocorpora.ru

carefully control the entire process, and be certain about the achieved result. Although we benefit from the existence of a highly advanced Komi morphological analyser (Rueter, 2000), we believe that this approach could also be applicable to other language documentation projects. In principle the analyser used for such projects would only need a very small lexicon containing those items to be pseudonymised, which could be very easily connected to the project's internal metadata.

2 Problem description

Much has been written on the problems with regard to the role of the linguistic consultants in language documentation, especially with regard to ethics and the acknowledgement of their role (see e.g. Rice, 2006; O'Meara and Good, 2010; Chelliah and Willem, 2010, 139–159; Dobrin and Berson, 2011; Bowern, 2015, 171–175). This may refer to making their identity known or not in publications, corpora and other sources, though acknowledging their role in the material collected, whether or not they wish to be overtly acknowledged by name, should in any case be done. As our aim is to share material as openly as possible we avoided collecting sensitive and personal information that could potentially be harmful for the individuals and communities when building our documentary corpus, and focused primarily on narratives that document local culture and history. Unfortunately, the current interpretations of EU legislation still leave some unclarity as to how questions of personal data in our recordings should be addressed.

Our research has been carried out in close cooperation with Zyrian Komi communities and native organisations, and we have made significant effort to ensure that our work is both accepted by the community and that the relevant materials are also available to community members. However, as e.g. Dorian (2010, 181) points out, consultants may not always fully grasp what exactly linguists plan on doing with the material they collect. Thereby we have to consider our own responsibility, independent of the informed consent provided by our language informants. We have to ensure that the ways the material is released to the public are appropriate.

The problem may be summarised such that even though we see no issues at the moment in sharing the complete dataset with the community, which we have already done in various ways (whilst taking into account the community's needs), and will also share it with researchers, most likely through a permission request with a description of intended use, it is currently not possible to share it entirely openly with the general public, as this would permanently expose community members' personal information. However, this kind of very formal and restricted method of distribution certainly hinders the active research use of the corpus, which is also something we do not want to happen. Thereby providing a pseudonymised version for the research use seems like a good alternative and something worth investigating further. When the pseudonymised version is available after an academically affiliated login in Korp, it is simple to familiarise oneself with the corpus to decide whether the complete dataset is needed for planned research. This also minimises the unnecessary redistribution of the entire corpus, as individuals do not need to access the complete dataset to evaluate its usability. In the same vein, with this information we can also create derived datasets that can be used in different experiments, but contain only minimal amount of personal data.

3 Method

The semantic tags that associate proper nouns in the Komi morphological analyser are: `Sem/Mal` (for a male forename), `Sem/Fem` (female forename), `Sem/Patr-Mal`, `Sem/Patr-Fem` (patronym), `Sem/Sur`, `Sem/Sur-Mal`, `Sem/Sur-Fem` (surname) and `Sem/Plc` (toponym). In addition to proper nouns numerals need specific attention, especially in constructions that are dates or years. Potentially, all tokens in the corpus tagged for one of these semantic categories contain information that either directly reveals the identity of individuals or information that can be easily used for revealing the identity of individuals when combined in combination with other data. The relevant identities can concern the recorded speaker(s) themselves (for instance the own name, names of parents and other relatives, or the place or date of birth) or they concern other individuals to which the recorded speaker is referring to. At the same time, however, spoken recordings contain names of places that are so large and general that they can be mentioned without in fact conveying a great deal of personal information. For example, cities such as Syktyvkar and Moscow have populations large enough such that identifying a person is usually not possible. The same is true for large bodies of water such as the rivers Izhma or Pechora, or mountain ranges

like Ural mountains – they span so many localities that they do not actually identify any of them. All these larger entities are treated through the method presented in Section 3.1. However, for very small settlements it is important to be more careful, as some locations only consist of individual houses, and speakers can thereby be identified more easily than would usually be the case. Example 1, which is taken from the introductory part of a personal interview, clearly illustrates how an individual utterance can contain different types of identifiable personal information, of which some are larger entities, i.e. the capital of the Komi Republic Syktyvkar, that can remain unmodified.

(1) Менӧ шуӧны Александр, ме ола Вертепын, велӧдчи Сыктывкарын.

menə	*ʃu-ənɨ*	*aʌeksandr,*	*me*
1SG.ACC	call-3PL.PRS	Aleksander	1SG
_	_	Sem/Mal	_
ol-a	*vertep-ɨn,*	*veləʨ:-i*	
live-1SG.PRS	Vertep-INE	study-1SG.PST	
_	Sem/Plc	_	
sɨktɨvkar-ɨn			
Syktyvkar-INE			
Sem/Plc			

'My name is Aleksander, I live in Vertep, I studied in Syktyvkar.'

We can therefore construct a list of major settlements that occur in our corpus, and allow those to pass unchanged through our pseudonymisation system. This is necessary, as the analyser would otherwise mark all places with the tag `Sem/Plc`. With regard to smaller localities, however, we have two options: 1) either mark them with an empty placeholder, or 2) replace the value with a specific "standard village". At the moment we have opted to use empty placeholders, although there are various options that should be considered. Example 1 illustrates how, using this logic, we can distinguish large localities of the type Syktyvkar from small ones such as Vertep.

Since the full name of each person we have worked with is included in our metadata database, it has been easy to evaluate whether all names are present in the Komi FST. Similarly, our metadata includes all names of recording locations, places of residence and places of birth. In practice, however, there are more place names mentioned in the narratives than those present in the metadata database, as the information in the database has originally been collected from those same interviews that were recorded and transcribed, i.e. the metadata referring to e.g. place names is limited to actual locations where people were born or lived or where recordings where made; place names merely mentioned in speech (like 'I visited Bangkok') have not been specially listed anywhere in our material. The work presented here is in principle one path toward constructing a more structured database of locations present in the corpus, which could be of high relevance for various types of linguistic and non-linguistic research using our data.

Another data type that potentially contains information sensitive to personal identities consists of dates. In the case of small local populations this can be true even for incomplete dates, i.e. indicating only the month of a year or even the year alone. Example 2 illustrates how this kind of information could be present in the corpus.

There is a tendency for such numbers to occur in formulaic expressions, especially as direct replies to questions about the age and such properties. In this kind of situation, when the numbers are pronounced in a very literary manner and are all in Komi, it is relatively easy to identify such segments. For instance, we can write a CG rule that targets sequences that contain the word for 'year' and a preceding sequence of numerals. Another way to target these segments would be to look into them as replies to questions where this content is asked. This would take advantage of the conversationality of the recordings.

However, there are particular challenges in those instances where the numbers are non-standard or in Russian, such as *пятого* 'fifth' in Example 2:

(2) Но ме рӧдитчи пятого декабря сюрс ӧкмыссё квайтымын витед воын.

no	*me*	*rəditʂ-i*	*pʲatovo*
well	1SG	be_born-1SG.PST	5th
_	_	_	Num/Card
dʲekabrʲa	*ʂurʂ*	*əkmisʂo*	
December	1000	900	
_	Num/Ord	Num/Ord	
kvajtɨmɨn	*vit-ed*	*vo-ɨn*	
60	5-CARD	year-INE	
Num/Ord	Num/Card	_	

'Well, I was born on the 5.12.1965'

The problem here is essentially that parts of the sentence are in Russian. Therefore, we cannot analyse it with the Komi analyser alone. The method has not yet been fully implemented for multilingual data, as processing such material contains numerous mixed forms and other problems. The process currently planned is to pass all unrecognised words through a Russian analyser, which, however, would also demand some consistency in tagging schemes used across these analysers. Cross-linguistic annotation schemes cannot always be straightforwardly matched (see discussion in Rueter and Partanen, 2019), but for a number of tasks any improvement here is very beneficial.

3.1 Implementation logic

Since CG does not allow us to modify the transcribed words directly, our method has been implemented through CG rules that add additional tags and prefixes to the available FST readings. For example, basic semantic tags do not need to be edited at this point, with the exception of locations that we want to keep, as described above. The CG rule that adds an additional tag for locations sufficiently large to be kept intact is very simple:

Keeping large toponyms

```
SUBSTITUTE:keep-large-places
(Sem/Plc) (Sem/LargePlc)
TARGET LARGE-PLACES ;
```

When this rule is applied, the later processing steps do not apply to locations marked with the *Sem/LargePlc* tag.

This rule depends on the list LARGE-PLACES which holds information about all the larger towns and settlements that we want to retain in the pseudonymised corpus. The list is manually compiled and contains some tens of generic large locations in Russia and elsewhere, among them common holiday destinations. It could be possible, however, to also connect this list to common open databases such as Wikidata,[3] in order to let the rule automatically apply to all settlements that have a population, for instance, over half a million. This, however, would move from the current direction where the rules are edited based on our own observations and thorough knowledge of the material, although the changes are implemented through the CG.

For removing birthday data we have experimented with a set of rules that are specific for that context. The rule **Explicit years** below briefly illustrates this logic, although the actual implementation is slightly more complicated with more word order variation included.

Explicit years

```
ADD:find-dates-years (Date)
TARGET (Num) OR (Ord)
((1* ("во")) OR (1* ("год"))) ;

ADD:find-dates-born (DateBirth)
TARGET (Num) OR (Ord)
((-1* ("чужны"))
OR (-1* ("рӧдитчывны"))
OR (-1*("рӧдитчыны"))) ;
```

Such contextual rules are useful as we can be relatively sure that this date is a date of birth, which is then tagged accordingly. There are, however, so many instances of dates that are without contiguous context that we have decided to use a rule that removes all years and dates. However, having the explicit information available about possible dates of birth in the corpus is very important and increases the accountability of the corpus creators.

All in all the system is relatively simple, consisting of some tens of CG rules. The actual removal of the sensitive tokens is done by a script reading the tagset that is specifically inserted through our rules. The script removes the actual tokens while leaving in the resulting pseudonymised corpus a placeholder-token, including the belonging morphosyntactic tags coming from the FST analyser.

The process is implemented as Python functions that are currently used as a part in the script pipeline that convert from the original corpus data in XML format used by ELAN into VRT format needed by Korp. There is, however, no reason why the same methods could not be adapted to other environments not working with ELAN or Korp.

The whole pipeline has been published in Zenodo (Partanen, 2019) and GitHub[4].

4 Evaluation

The quality of the system was evaluated with one pass through the complete corpus, and another more qualitative examination of one individual recording

[3] https://www.wikidata.org

[4] https://github.com/langdoc/langdoc-pseudonymization

of five speakers. This gives a relatively good impression of the accuracy and also the usability of the method. If the resulting text is unusable with too many omitted sections it is clear that the method is not of particular use for researchers.

In the complete corpus currently 2% of all tokens get marked as being possible proper names of persons, places or dates. During evaluation we also found that it was necessary to adjust the system so that both Russian and Komi language versions of the names of settlements are included in the analyser, as the speakers may use both. Some toponyms for smaller places, for instance *Мохча*, were missing from the analyser, as were some less-used patronyms, for instance *Парфёнович* and *Арсентьевна*. Adding these is, obviously, a trivial task. However, looking also at the Russian analyser benefits the infrastructure at large, as these same names occur in various languages spoken in Russia.

Interestingly, our evaluation run revealed also situations where the system *en passant* removes ambiguously tagged content. One such example is the lemma *Бура*, which could be analysed as a surname or as an adverb. However, to our knowledge it is only used as an adverb with the meaning 'well' and never as a surname. The problem is related to various names that are foreign in cultural context in Komi, but in theory could be foreign names. Also several common Russian words have a potential surname reading, which by our evaluation is not relevant in our corpus. These are, for example, *Горячий*, *Ден* and *Готов*. We have relaxed the the system with additional rules to ignore such cases, but with careful consideration only.

One culturally important feature of our method is that it can correctly detect native Komi names, such as the multi-word *Пась Коля*. Among surnames a category of individuals who are so well known that they do not need to be removed is still under consideration. With some names this is clear, for instance, all tokens *Вихман* 'Wichmann' occurring in the corpus refer to the Finnish researcher Yrjö Wichmann. Similarly, names such as 'Jesus' or 'Lenin' are kept, as these names are not used as a given name or nick name in our cultural context. Other surnames, such as *Лыткин* 'Lytkin', may either refer to the well known Komi researcher Vasily Lytkin or other persons with the same name. Therefore, this part of the system needs refinement.

One benefit of using an orthographic transcription system for a spoken-language corpus (instead of phonemic transcription, cf. Blokland et al., 2015; Gerstenberger et al., 2016) is that orthographically proper nouns are consistently written with initial uppercase. Their parsing and verification for our pseudonymisation system is therefore an easy process. Our evaluation run revealed a list of approximately 7000 tokens with initial uppercase letters, that were not being recognised by the analyser yet. It has been possible to go through this list manually while collecting the forms. As these were primarily items missing from the lexicon, their inclusion was simple. Note also that since our field recordings have been carried out in a limited number of communities, the same toponyms are repeated in different recordings. All work with including these missing names in the analyser's lexicon files rapidly improves the performance of the analyser overall. All in all, our examination resulted in approximately 250 new lemmas being added into the lexicon files of the Komi morphological analyser.

In one particular category of toponyms, names derive from common nouns designating landscape features, such as *Ди* /di/ 'island' and *Ёль* /joʎ/ 'stream'. In such instances the pseudonymisation is done only when they are written with capital letter and are in singular. Although this rule overgeneralises a few relevant common nouns in sentence initial position, it seems to handle this problem sufficiently. Some concepts are, however, so generic that to our knowledge they do not refer uniquely to individual settlements specifically, i.e. *Яг* /jag/ 'forest´, *Курья* /kurja/ 'bay´ and *Катыд* /katɨd/ 'downstream´. The proper noun reading is left at place when the form is written in lower case, but does not have any other possible interpretations. This is against our transcription conventions, but it is beneficial that the system has some robustness for such instances where the spelling is by mistake deviating from our guidelines.

The evaluation of all tokens marked in the entire corpus as potentially containing personal information revealed that out of 8000 pseudonymised tokens only 4% were mistakenly removed. If we would had pseudonymised all items, which are semantically tagged as proper names or dates, without implementing further rules as described above, the ratio of mistakenly removed forms would had been almost 50% of the tagged ones. This is so primarily because very high frequency words such as *Из* /iz/ 'stone; Ural mountain; negation verb form' and *Коми* /komi/ 'Komi (Republic); Komi (an

ethnic group, a language)' would had been tagged for pseudonymisation and removed by the script as well. This illustrates well, that such a task as writing rules for pseudonymisation can only be done with careful understanding of the data and its cultural context. The forms that cannot be proper nouns under any circumstances have been added to a separate list. There is always the possibility to edit such instances also at the level of the analyser itself.

5 Conclusion & Further work

Later evaluation should be linked to explicit tests that demonstrate that the rules are working under the desired conditions. However, already now the pipeline proposed in this paper has proven itself to be highly effective for the pseudonymisation of a large spoken-language corpus resulting from fieldwork recordings of an endangered language. The specific merits of this system are that it is easy to extend, and through rule-based implementation its precision can be very reliably evaluated and adjusted.

One possible, and already planned, utilization for our method is the selection of sentences that can be included into dictionaries as examples. There is a general need to display in different web interfaces spoken language sentences that illustrate how a word is used, and through our method we could automatise the task to select example sentences. This could be combined into modern dictionary interfaces such as those discussed by Rueter et al. (2017) and Hämäläinen and Rueter (2018).

We have described a method that is effective, reliable, and meets a concrete need in corpus data processing. We have also presented a novel and creative application for Constraint Grammar. We want to stress that besides removing or editing the marked personal information this method could also be used to evaluate how much of this kind of information individual transcriptions contain, thereby providing rough metrics about their level of sensitivity.

Since no anonymization or pseudonymisation method is perfectly reliable, the materials cannot be made entirely available without further manual verification. We believe, however, that the results we have achieved reach a level that does allow relatively open distribution with only basic user authentication, for example, within an academic research context. As the system described contains many changing elements: Komi FST, CG and the corpus itself, testing and refinement will necessarily continue.

Acknowledgments

The authors of this paper collaborate within the project "Language Documentation meets Language Technology: The Next Step in the Description of Komi", funded by the Kone Foundation, Finland. Special thanks to the University of Helsinki for funding Niko Partanen's travel.

Thanks to Jack Rueter for his valuable and continuous work on the computational description of Komi, and to Marina Fedina's team at FU-Lab in Syktyvkar for their ongoing efforts in building Komi language resources and language technology.

References

Malin Ahlberg, Lars Borin, Markus Forsberg, Martin Hammarstedt, Leif-Jöran Olsson, Olof Olsson, Johan Roxendal, and Jonatan Uppström. 2013. Korp and Karp–a bestiary of language resources: the research infrastructure of Språkbanken. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 429–433.

Eckhard Bick and Tino Didriksen. 2015. Cg-3—beyond classical constraint grammar. In *Proceedings of the 20th Nordic conference of computational linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*, 109, pages 31–39. Linköping University Electronic Press.

Rogier Blokland, Marina Fedina, Niko Partanen, and Michael Rießler. 2020. Spoken Komi Corpus. The Language Bank of Finland version.

Rogier Blokland, 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*, number 2015:2 in Septentrio Conference Series, pages 8–18. The University Library of Tromsø.

Claire Bowern. 2015. *Linguistic fieldwork: A practical guide*. Springer.

Shobhana L Chelliah and Jules Willem. 2010. *Handbook of descriptive linguistic fieldwork*. Springer Science & Business Media.

Lise M Dobrin and Joshua Berson. 2011. Speakers and language documentation. In *The Cambridge handbook of endangered languages*, pages 187–211. Cambridge University Press.

Nancy C Dorian. 2010. Documentation and responsibility. *Language & Communication*, 30(3):179–185.

Ciprian Gerstenberger, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2016. Utilizing language technology in the documentation of endangered Uralic languages. 4:29–47.

Mika Hämäläinen and Jack Rueter. 2018. Advances in synchronized XML-MediaWiki dictionary development in the context of endangered Uralic languages. In *Proceedings of the eighteenth EURALEX international congress*, pages 967–978.

Nikolaus Himmelmann. 2006. Language documentation. 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.

Mika Hämäläinen. 2019. UralicNLP: An NLP library for Uralic languages. *Journal of Open Source Software*, 4(37):1345.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing running text. In *COLNG 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*.

Fred Karlsson, Atro Voutilainen, Juha Heikkilae, and Arto Anttila. 2011. *Constraint Grammar: a language-independent system for parsing unrestricted text*, volume 4. Walter de Gruyter.

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2009. Hfst tools for morphology–an efficient open-source package for construction of morphological analyzers. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 28–47. Springer.

Carolyn O'Meara and Jeff Good. 2010. Ethical issues in legacy language resources. *Language & Communication*, 30(3):162–170.

Niko Partanen. 2019. langdoc/langdoc-pseudonymization: Language documentation corpus pseudonymization method.

Keren Rice. 2006. Ethical issues in linguistic fieldwork: An overview. *Journal of Academic Ethics*, 4(1-4):123–155.

Jack Rueter and Niko Partanen. 2019. Survey of Uralic Universal Dependencies development. In *Workshop on Universal Dependencies*, page 78. Association for Computational Linguistics.

Jack M. Rueter. 2000. Хельсинкиса университетын кыв туялысь Ижкарын перымса симпозиум вылын лыддьӧмтор. In *Пермистика 6 (Proceedings of Permistika 6 conference)*, pages 154–158.

Jack Michael Rueter, Mika Kalevi Hämäläinen, et al. 2017. Synchronized Mediawiki based analyzer dictionary development. In *3rd International Workshop for Computational Linguistics of Uralic Languages Proceedings of the Workshop*. Association for Computational Linguistics.

Mandana Seyfeddinipur, Felix Ameka, Lissant Bolton, Jonathan Blumtritt, Brian Carpenter, Hilaria Cruz, Sebastian Drude, Patience L Epps, Vera Ferreira, Ana Vilacy Galucio, et al. 2019. Public access to research data in language documentation: Challenges and possible strategies. *Language Documentation and Conservation*, 13:545–563.