Using computational approaches to integrate endangered language legacy data into documentation corpora: Past experiences and challenges ahead

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

Niko Partanen
Institute for the Languages of Finland
niko.partanen@kotus.fi

Michael Rießler
University of Bielefeld
michael.riessler@uni-bielefeld.de

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

Abstract

The systematic integration of pre-digital published transcriptions of legacy language materials offers many possibilities to enrich documentary corpora with data that is often very comparable to contemporary collections, and often originating from the same speech communities researchers currently work with. Especially recent advances in text recognition technologies make the reuse of old materials a very attractive and accessible task. However, the output of text recognition needs to be connected to further parts of the pipeline, namely forced alignment and speech recognition. The workflows discussed here attempt to reach a maximally useful situation where legacy data is transformed into a usable and comparable format, but not yet transformed into a time aligned corpus.

1 Introduction

This paper discusses opportunities for and challenges of an approach in documentary linguistics which systematically integrates previously published, pre-digital heritage data into a corpus. Based on our own experience, we aim to develop better practices and standards for building more significant corpora in the context of endangered language documentation and description, including potentially any available linguistic data beyond our own annotated fieldwork recordings. Corpora are not only needed for empirically sound descriptions of endangered languages, but can also be utilized in various ways in future computational linguistic studies on these languages.

Although in many cases language documentation work starts from scratch, this is not always the case, such as when previous generations of researchers have produced very large recorded and transcribed collections for the same languages, sometimes even with the same language communities or ancestors of the current speakers. Woodbury (2003) mentions the curation of huge tape collections as upcoming work, and since these collections often connect into already transcribed and published versions of the given texts, our approach aligns very closely with this task. The relevant materials may be handwritten, or partially published in print, and the original recordings are usually scattered in various archives and personal collections, possibly forgotten or even lost. Including heritage data in contemporary corpora is not an easy task. Indeed, it can be overwhelming and very challenging, which makes the temptation to work primarily with new, self-collected data very strong. However, we argue that heritage data is important enough that resources should be systematically allocated to including these data in language documentation projects.

Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages: Vol. 2 Extended Abstracts, pages 24–30,
Honolulu, Hawai‘i, USA, February 26–27, 2019.

The authors of this paper have worked extensively with Zyrian Komi, and various Saamic languages (all in the Uralic language family). Examples of the publications integrated into our corpora are parts of Yrjö Wichmann's *Syrjänische Volksdichtung* (Komi spoken texts collected in 1880s, published in 1916), T.E. Uotila's *Syrjänische Texte* (Komi spoken texts collected in 1940s, published in 1986–2006) and Erik Vászolyi's *Syrjaenica* series (Komi materials collected in 1960s, published i.e. in 1999), Arvid Genetz' *Sprachproben* (Akkala, Kildin, Skolt, and Ter Saami spoken texts collected in the 1870s, published in 1891), Georgi Kert's *Obrazcy saamskoj reči* (Kildin and Ter Saami spoken texts collected in the 1950s and 1960s, published in 1961), as well as Ignácz Halász' Pite Saami text collections published in 1893, Eliel Lagercrantz' Pite Saami texts from 1921 (published in 1957 and 1963) and numerous archived materials collected by Israel Ruong throughout his carrier.

Whereas the oldest materials, such as those by Wichmann and Genetz, are already in the Public Domain, the reuse of newer materials had to be negotiated with different stakeholders if the data has not been openly licensed by the publisher already. Methods for accessing these materials have included manual retyping, retrieving text from original digital files and building new OCR models for digitization. Although there may be a time and place for such approaches, this paper emphasizes the most automatized methods, with the wish to streamline the process even further.

Our discussion focuses on text collections published for scientific use, typically as aligned transcriptions and translations in a monograph. This is somewhat distinct from the needs that arise around the use of other community-created resources, such as literature and other truly written-mode texts. Another topic that we do not discuss here is the digitization of dictionaries (as addressed e.g. by Maxwell and Bills (2017)). We do not focus on specific technical implementations, as these change quickly, but discuss the topic on a more conceptual level. However, our technical pipeline has been documented in a GitHub repository[1] and is openly available.

2 Methodological background

One distinct feature of our work, compared to common methodology in fieldwork-based language documentation projects, has been the continual application of language technology in corpus annotation (Blokland et al., 2015; Gerstenberger et al., 2016, 2017). Using computational linguistic approaches for more automated corpus annotation, a component of Documentary Linguistics which was not mentioned by Himmelmann (1998), has resulted in relatively large corpora (measured in the number of morphosyntactically tagged tokens). Furthermore, we consistently integrate all available legacy data, in addition to our own fieldwork recordings. This is possible because the endangered Northern Eurasian languages we work on have a long research tradition and possess a number of extant textual sources in addition to preliminary descriptions. Text collections published by pre-digital language documenters since the late 19th century century are especially interesting for language documentation. The similarity to contemporary language documentation materials may not be immediately obvious, because typically no audio representations of these texts exist (if they predate contemporary recording technology), or audio representations are not available together with the original recording (if the original recording was archived but not catalogued properly or not archived at all). However, the texts correspond to recent transcribed recordings because they represent transcribed spoken linguistic events; furthermore, they are often accompanied by translations into majority languages, just as in modern language documentation projects. The lack of interlinear glossing does not necessarily differentiate these materials from contemporary work, as the need for such annotations is well worth questioning anyway from a documentation perspective when dealing with languages for which basic phonological and morphological descriptions are already available (like for most endangered languages of Northern Eurasia).

Our practices have concentrated around

[1] github.com/langdoc/ocr-pipeline

efforts to digitize these materials and turn them into structured corpora using the quasi-standard ELAN xml-format.[2] We use ELAN even when no recording exists or none has been made available. This is done in order to restrict the data carrier formats of our corpora to a single format (ELAN); due to the technical requirements of ELAN, utterances are symbolically ”time-aligned” in each ELAN file, although time-alignment is irrelevant for such exclusively written-format heritage texts.[3] This solution arises primarily out of convenience. For interoperability with audiovisual materials in documentary corpora and in order to query the whole corpus effectively, we want a systematic structure across all corpus files. Our projects have evolved from fieldwork-based language documentation and ELAN is the best-suited tool we have encountered for aligning audio (and video) recordings with annotations. ELAN also allows offline corpus searches[4] and it has become a quasi-standard for archiving language documentation data.

We keep all original transcription systems as separate tiers in the resulting corpus, while using the primary transcription tier in the same orthographic representation relevant for each of our contemporary documentation projects. As long as the interpretation of the corresponding phonological system is similar, the transliteration from one system to another is relatively easy; however, doing this consistently and most reliably is still a question that needs more attention. With legacy data it is not unusual for each publication to use a different transcription system, so many conversion patterns are needed. Our projects use Git to ensure version control of the ELAN files; this solves the issue to some extent, but still needs additional conventions to keep track of the actual modifications. All computational methods we have used to transliterate between writing systems have been entirely rule-based, and we are currently developing a Finite State Transducer for this.

3 Recent advances in text recognition

Whether there are enough digital texts available or whether Optical Character Recognition (OCR) tools are needed extensively varies from case to case (cf. Arppe et al., 2016, 5). Working with Uralic languages, publications since the 1980s can occasionally be found as original digital files, which, although coming with a myriad of other problems, are usually the easiest source for text retrieval. In cases where the text collections were typeset by hand or when the original text files were lost, the only solutions are retyping the text or performing text recognition.

Thus far it has been challenging to carry out good quality OCR on complex scripts with a large number of diacritics, although we have had minor success with commercial software such as Abbyy FineReader, a solution which comes with additional issues (Partanen, 2017). Good results with open source software have previously been tied to the availability of matching fonts, which used to be a great impediment. In recent years, open source OCR systems such as Tesseract[5] and Ocropy[6] have shifted towards neural network approaches that process individual lines instead of characters; this streamlines the training process. It is still difficult to train OCR models that work consistently across texts from a variety of sources, yet training a model for the transcription used in a single publication or one writing system seems to be doable with surprisingly little effort. In the tests done by Partanen and Rießler (2019), a few hundred lines of manually created training data were enough to bootstrap a useful OCR system.

Many text collections for endangered languages can be considered representative of a complex but very narrow domain for text recognition purposes, but here the fact that machine learning methods tend to excel in tasks with such conditions is a clear advantage. If a specific transcription system is used only in one publication series, there is no need to recognize this writing anywhere outside that specific publication, so we can easily train an OCR system that only works within this con-

[2] `tla.mpi.nl/tools/tla-tools/elan/`

[3] If audio becomes available, it can be added and the annotations aligned later.

[4] Note however that the corpus search capabilities of the program have some questions that need to be addressed, see Wilbur 2019

[5] `github.com/tesseract-ocr/tesseract`

[6] `github.com/tmbdev/ocropy`

text. The aforementioned study by Partanen and Rießler (2019) tried this approach successfully with texts on different languages which used the same writing system. The results indicate that it is possible to train multilingual OCR system as long as the typeface is identical and all characters are within the training data. These approaches could well be extended to publications using the International Phonetic Alphabet, the Uralic Phonetic Alphabet or others.

On a related note, methods for Handwritten Text Recognition (HTR) have also been rapidly improving (Kahle et al., 2017). Instead of being font- or typeface-specific, HTR systems generally learn a specific person's handwriting style with considerable accuracy. However, HTR currently needs more training data than OCR (hundreds of pages to achieve ideal results), and this restricts its application to situations where the necessary data exists. In order to test HTR tools with legacy transcriptions, the Institute for the Languages of Finland (KOTUS) is currently carrying out a series of experiments using the Transkribus platform[7] to recognize handwritten transcriptions of dialectal Finnish. In this case there are approximately 17,000 transcribed pages produced by one person, Eeva Yli-Luukko, between the 1960s and 1980s. A few hundred pages are now aligned line-by-line, and this seems to be enough to reach a recognition accuracy higher than 90%. M.A. Castrén's handwritten notes on endangered Siberian languages from the 19th century run to over 10,000 pages, and preliminary experiments carried out with this material at KOTUS have reached a recognition accuracy of 75%. These numbers are still far from the results achieved with state-of-the-art OCR systems (which reach over 99% accuracy), but further development and fast progress on this front is sure to happen. The HTR model training is based on material annotated manually in the Manuscripta Castréniana project, which also publishes digital editions.[8]

4 Towards structured data

In a typical language documentation corpus, the transcribed utterances are time-aligned to the recorded audio. The metadata contain additional information about the speech event and participants. Since the audio represents the primary data on which the transcription is based, the ELAN file itself stores the links between audio and transcription. In some situations with legacy texts, the recorded audio does not exist, and then the representation on the page can be considered the closest we have to the primary data. This raises the question whether instead of audio links we would need to store information about the sentence's coordinates on the page. Modern OCR software has XML export formats that contain this information, and technically it is possible to connect the coordinates to utterance references. This may seem unnecessary when the text comes from a more contemporary publication, but the older and rarer the source, the more essential this seems. This information concerning the location on the page is essential if one wants, for instance, to link an image to the page with highlighting, or to connect annotations to a digital facsimile of the publication. This linkage becomes more complex when there are both audio and printed images representing the same speech event, although technically the utterance specific metadata could be just enhanced with the time codes. However, this becomes even more complicated when the text and the audio deviate from one another, and the transcription is manually corrected and edited. In this setting, both what would be normally transcribed in ELAN and what has been published in the text collection are derivatives of the original audio. The links between these versions would be useful for various purposes, but this becomes difficult when the utterances are eventually edited and corrected in the ELAN version which has the audio link. Published versions are often edited in various ways, which leads to a mismatch between the published text and the original speech event.

Written text may contain inherent structures, such as information about who is speaking which utterance, and what the consecutive order of the utterances is. Usually no refer-

[7] transkribus.eu

[8] sgr.fi/manuscripta/

ence to time codes is made, so the text and audio have to be aligned separately. Forced alignment tools have been tested with language documentation data and even suggested for use with legacy data (Strunk et al., 2014, 3942). However, in our experience, the lack of exact correspondence between audio and transcription, combined with the frequent overlapping speech common in spoken language, prevents current forced aligners from performing sufficiently. Because this work – at least theoretically – should someday be possible, we have not engaged in extensive manual alignment, but are waiting for improved automatic tools to be developed.

Splitting utterances or word-length segments into words and phonemes already works with significant accuracy (cf. Kempton, 2017). Language documentation projects often produce enough transcriptions aligned with audio that some sort of a customized forced aligner could be trained using this data, as has been tested recently with Tongan documentary data (Johnson et al., 2018). Similarly, the experiments with speech recognition for endangered languages discussed in Foley et al. (2018); Adams et al. (2018) offer some new possibilities for forced alignment as well, since erroneously recognized speech could probably be matched with more correct textual transcriptions.

5 Wider perspectives and connections

Older publications that contain text collections can easily be found in bibliographical databases, but with archival records, it is more complicated. Archive identifiers can exist for transcriptions in older publications, although sometimes there is only a note indicating the materials exist. We have had occasional success finding recordings described in text collections through harvested archive metadata, but not all archives release their metadata this way (Thieberger, 2016); indeed, in our experience, even when metadata exists it may not be sufficient to identify which recording matches which text.

When the recordings can be located, acquiring the original recordings has been fairly straightforward, and ensuring more wide-ranging usage rights for integrating this material into a corpus can be negotiated together with the archive and the publisher. The majority of material we have worked with was originally collected by researchers who are now deceased, but in some cases the original author may still be alive and interested in participating in digitization efforts. However, note that the archives have traditionally obtained full rights to redistribute the material upon receiving the original recordings. That said, the copyright and ownership questions concerning this kind of resources can be complicated. Unarchived collections are likely the greatest source of difficulties in this respect.

The large number of texts contained in these published text collections alone is reason enough to integrate them without exception into language documentation materials. There are numerous benefits to having as large a corpus as possible, not least from the point of view of language technology. However, there are reasons that go beyond the simple utility of having a large corpus, the most significant of which is more human and connected to working with the speech communities. Typically these materials have been collected by the relatives and ancestors of community members, and sometimes even include speakers still alive today. It is not uncommon for these small print publications to have never been made accessible in the regions they originate from. There are also clear scientific interests in using such materials, and they can be used to plan future documentation work.

With all the languages we have worked with, some of the oldest materials used are in Public Domain, and one possibility which we have been investigating is the publication of the annotated portions of the corpora with very open licenses so that at least some part of the corpus could be used by researchers with no restrictions. This approach has recently been continued by including these resources in treebanks within a Universal Dependencies project (Partanen et al., 2018). However, we are still lacking methodology and practices that would allow us to seamlessly combine distinct research outputs, such as treebanks, into primary materials and archived versions, so that discoverability of all components would be ensured, even when the data changes and different resources are stored in different repositories: scanned pages in a library's digital system, original audio recordings in one archive, the partly derived language documentation corpus in another archive, and the treebanks in their own repository.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of LREC 2018*.

Antti Arppe, Jordan Lachler, Trond Trosterud, Lene Antonsen, and Sjur N Moshagen. 2016. Basic language resource kits for endangered languages: A case study of Plains Cree. *CCURL*, pages 1–8.

Rogier Blokland, Ciprian Gerstenberger, Marina Fedina, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2015. 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*, number 2015:2 in Septentrio Conference Series, pages 8–18. The University Library of Tromsø, Tromsø.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T Mark Ellison, Daan van Esch, Scott Heath, František Kratochvíl, Zara Maxwell-Smith, and David Nash. 2018. Building speech recognition systems for language documentation: The CoEDL endangered language pipeline and inference system (ELPIS). In *6th Internationall Workshop on Spoken Language Technologies for Under-Resourced Languages*.

Arvid Genetz. 1891. *Wörterbuch der Kola-Lappischen Dialekte nebst Sprachproben*. Number 50 in Bidrag till kännedom af Finlands natur och folk. Finska Vetenskaps-Societeten, Helsingfors.

Ciprian Gerstenberger, Niko Partanen, and Michael Rießler. 2017. Instant annotations in ELAN corpora of spoken and written Komi, an endangered language of the Barents Sea region. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 57–66. Association for Computational Linguistics, Honolulu.

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

Ignácz Halász. 1893. *Népköltési gyűjtemény*, volume 5 of *Svéd-Lapp Nyelv*. Magyar tudományos akadémia, Budapest.

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

Lisa M Johnson, Marianna Di Paolo, and Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing prosodylab-aligner with tongan data. *Language Documentation and Description*, 12:80–123.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In *International Conference on Document Analysis and Recognition (ICDAR 2017)*, volume 4, pages 19–24.

Timothy Kempton. 2017. Cross-language forced alignment to assist community-based linguistics for low resource languages. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 165–169, Honolulu. Association for Computational Linguistics.

Georgij M. Kert. 1961. *Obrazcy saamskoj reči. Materialy po jazyku i fol'kloru saamov Kol'skogo poluostrova (kil'dinskij i iokan'gskij dialekty)*. Nauka, Moskva.

Eliel Lagercrantz. 1957. West- und südlappische Texte. Gesammelt und herausgegeben von Eliel Lagercrantz. In (Lagercrantz, 1957–1966).

Eliel Lagercrantz. 1957–1966. *Lappische Volksdichtung*. Number 112,115,117,120,124,126,141 in Mémoires de la Société Finno-ougrienne. Finno-Ugrian Society, Helsinki.

Eliel Lagercrantz. 1963. Texte aus den see-, nord-, west- und südlappischen dialekten: gesammelt, übersetzt und herausgegeben von Eliel Lagercrantz: Index: Verzeichnis der motive und varianten: mythische symbolwelt: Stilkunst und sprache. In (Lagercrantz, 1957–1966).

Michael Maxwell and Aric Bills. 2017. Endangered data for endangered languages: Digitizing print dictionaries. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 85–91.

Niko Partanen. 2017. Challenges in OCR today. report on experiences from INEL. In *Electronic Writing of RF Peoples: History, Issues, and Perspectives. 16.–17.3.2017, Syktyvkar*, pages 263–273.

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first Komi-Zyrian Universal Dependencies treebanks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 126–132. Association for Computational Linguistics, Brussels.

Niko Partanen and Michael Rießler. 2019. An OCR system for the Unified Northern Alphabet. In *International Workshop on Computational Linguistics for Uralic languages (IWCLUL 2019)*. Association for Computational Linguistics, Tartu.

Jan Strunk, Florian Schiel, Frank Seifart, et al. 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using webmaus. In *International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3940–3947, Reykjavík.

Nick Thieberger. 2016. What remains to be done—exposing invisible collections in the other 7,000 languages and why it is a DH enterprise. *Digital Scholarship in the Humanities*, 32(2):423–434.

Toivo Emil Uotila. 1986–2006. *Syrjänische Texte*. Number 186,193,202,221,252 in Mémoires de la Société Finno-ougrienne. Finno-Ugrian Society, Helsinki.

E. Vászolyi-Vasse. 1999. *Syrjaenica. Narratives, Folklore and Folk Poetry from eight dialects of Komi. Upper Izhma, Lower Ob, Kanin Peninsula, Upper Jusva, Middle Inva, Udora*, volume 1 of *Specimina Sibirica*. Seminar für Uralische Philologie der Berzsenyi Hochschule, Szombathely.

Yrjö Wichmann. 1916. *Syrjänische Volksdichtung*, volume 38 of *Mémoires de la Société Finno-ougrienne*. Finno-Ugrian Society, Helsinki.

Joshua Wilbur. 2019. ELAN as a search engine for hierarchically structured, tagged corpora. In *International Workshop on Computational Linguistics for Uralic languages (IWCLUL 2019)*, Tartu. Association for Computational Linguistics.

Anthony C. Woodbury. 2003. Defining documentary linguistics. In Peter K. Austin, editor, *Language Documentation and Description*, volume 1, pages 35–51. SOAS, University of London, London.