Challenges in OCR today: INEL experiences

Niko Partanen

31 January 2017

Contents

Introduction

This paper discusses some experiences and workflows from INEL project (*"Grammatiken, Korpora und Sprachtechnologie für Indigene nordeurasische Sprachen"*)(https://inel.corpora.uni-hamburg.de). The long-term project is funded within the framework of the Academies' Programme, which is coordinated by the Union of the German Academies of Sciences and Humanities.

In order to preserve the written cultural heritage, digital preservation is necessary, and there are already good standards for digitalization of books and manuscripts. The archiving projects can easily justify concentrating only to preservation of legacy materials, referring with this mainly to digitalization and metadata related catalogization. The next logical step after preservation is to bring these materials into common and scientific use. Opening this kind of datasets is still in early phase, and many related practices are largely underdeveloped. For example, in Scandinavia the libraries have recently offered digitalized materials to the public use, although the focus has been in publications before 1910 (Pääkkönen et al. 2016). Within language documentation projects the concept of open data is barely discussed, and the most of the data seems to be behind differently restricted licenses or then with limitations to only academic use. In many cases this is needed due the personal privacy issues which are related to the nature of the data collected from private individuals. However, in part this must be connected to lack of knowledge and good practices with the licensing and data sharing more generally. This is very unfortunate, since incorrectly structured and licensed materials will take considerable future effort and negotiation to get them into active research use.

In order to do provide better usability the content has to be accessible, and with the written documents this refers mainly to texts. Optical Character Recognition (henceforth OCR) tools are irreplaceable in efforts to bring printed and hand written materials into further use. OCR aims to produce a *text encoding* from the image data. This refers explicitly to detection of individual signs in the text, which can be represented in the resulting document in different ways. The current preferred standard would be representation with closely matching Unicode characters, in case those are available. The coverage of Unicode has extended continuously, but there are still few dark corners left. However, it is clear that part of the content is in the structure of the text document, which means that simply retrieving the text is only one step. In my opinion it is still somewhat unclear what is the ideal format to store this restructured data so that it is most useful for linguists and other researchers. There are already available more wider and systematic comparisons of different OCR tools, see for example Tafti et. al. (2016), but I aim to contribute to the discussion by pointing out particular needs of language documentation, which possibly are relevant for the wider linguistic community as well.

During the last years there have been multiple very important and large digitalization projects, run both by academic institutions and libraries, often in cooperation. Some of these, such as Endangered Archives ran by the British Library, have focused mainly to audio and photograph mediums. Others, namely Fenno-Ugrica and Komi Kyv website maintained in Syktyvkar, are good examples from the systematic digitalization of written resources. This opens entirely new situation for the research on these languages, and also allows the speakers to access this content on their language in ways which have never before been possible. Since many projects have produced millions of scanned pages, the principal question is how we can use this data. By use I refer primarily to the use in scientific research and also would prioritize highly use which makes materials better available for the community of speakers themselves. As we are discussing mainly minority language materials, the commercial use is not the most relevant issue. Also from this perspective working primarily with data that can be released on permissive licenses is the most advisable approach. Unfortunately very few projects have adopted clear licenses on the data, mainly Fenno-Ugrica, which has decided to use Public Domain. It is foreseeable that in the future this will be one of the questions which have to be answered in relation to different uses.

In INEL project we have been working with the digitalization of Selkup, Dolgan and Kamas materials. In contrary to the projects mentioned earlier, our focus has been primarily in the published fieldwork materials and other representations of spoken primary data on these languages. Since INEL is a long term project, the workflows and conventions established now will be extended further into new languages. Komi-Permyak and Komi-Yazva will also be included in INEL, which is relevant in the Permic context of this conference as well. The texts we work with tend to fall into two distinct categories: hand written manuscripts and published printed transcriptions. This data is also unusually complex because the transcriptions are usually related into different archival audio recordings, but the relation may not be entirely transparent. Currently Daniel Jettka is implementing a graph database as a solution to store interlinked data, which looks like a solution with many good sides to deal with complicated data model inherent with our data. For example, with some Selkup transcriptions the recordings are apparently different versions from the ones recorded. At the same time the same transcriptions may have been published in several publications in somewhat different versions and using various transcription conventions. The same recordings may be stored in different formats and copies in various archives. All this brings additional layer of complexity to the metadata associated with the primary data.

The situation is similar with digitalization of fieldwork materials on majority of Uralic languages spoken in Russia. This is also a very particular case for OCR tools since the scripts used are often not included in available OCR models, the number of diacritics can make the scripts extremely complex and the fonts used may be very customized for specific publications. Indeed the variation between the publications may often be so high that it is necessary to question how much effort has to be put into OCR development, if the specific variety is used only on few pages in one publication. It seems that in many ways the OCR challenges with printed transcriptions could be very comparable to the OCR of printed early modern manuscripts, although there are also clear differences, possibly mainly in the volume of the texts (transcriptions tend to be relatively small collections in the end, with maybe 300 pages in one publication at maximum). In many cases also the copyright questions tend to get very complex with the legacy materials, since the rights of the speakers, original researchers, archives and publishers may all interact in many ways. That said, it seems we are still far away even from solving the copyright issues to a very simple data, so the question should not be made overly complicated in this point.

In INEL we have currently implemented OCR tools only to printed texts, but in principle to extend this into hand written texts will be in some point necessary. Since the hand written text recognition is still actively developing field, the choise of software on this domain is narrow. One good candidate to take into consideration is Transkribus, developed in Innsbruck, Austria, but very likely other solutions will still emerge. Also here it has to be taken into account that the distinction between hand written and printed is possibly less relevant than the distinction between combining and distinct characters.

Basics of OCR workflow

Most of the time the first step after digitalization is to find some solutions to perform OCR on the text. To take this further, there are several tasks which have to be undertaken before the data is actually ready to be used by general audience or researchers. It seems to me that many of these tasks have not been commonly discussed, which is why I want to present some of my thoughts on the topic in this paper. Although not strictly speaking new research, this paper aims to introduce several concepts into discussion when the OCR tools are evaluated and chosen. Within INEL project (Wagner-Nagy et al. 2015) we have done a wide ranging evaluation of different tools currently available for tasks relevant for us, and despite our satisfactory results one must also acknowledge the existence of several alarming bottlenecks, insuitable technical combinations and dead ends within the existing workflows.

Moreover, it must be recognized that mere OCR of the text is often not enough in order to provide materials ready for linguistic analysis, but further restructuring is necessary. Often it adds value to automatically extract more information than just the text, as many documents contain systematically information about the author, location and time, among many other metadata variables. Information about the position and layout for characters, lines and words on pages is also among such data. The need for this is of course entirely different depending from the text type, since it is entirely possible to represent a book as a linear text paragraphs, which is not possible for newspaper texts which are divided into more complex layout elements. As a more technical note, also the recognition confidence for individual characters could be very useful for further post processing. This naturally grows importance with languages which have less mature OCR environment. In this paper different problems are identified, and as far as possible, alternatives are offered and solutions envisioned.

Although often we can rely on command line tools, there are some tasks which are particularly unsuitable to be done with graphical user interface. As mentioned, one of these is post-processing of OCR result. Another is manual correction of the detected text regions, since their correct definition is necessary for succesful OCR. Both of these tasks in almost all cases demands visual comparison between the recognized text and the original image. There are plenty of different tools which offer possibility to manually correct OCR'd text, and virtually all of them have very similar user interfaces. The page image is on the left, and the text field on the right, and moving cursor around the text field has some kind of correspondence on top of the image. This makes one wonder why so many projects have ended up to the solution to develope new tools for post-correction. Fenno-Ugrica(https://github.com/NatLibFi/ocrui-frontend) has their own tool, Deutches Textarchiv(http://www.deutsches-textarchiv.de/doku/softwar) has their own. PoCoTo(http://thorstenv.github.io/PoCoTo. natalialevshina.com/corpus.html)(http://www.natalialevshina.com/corpus.html) by Natalia Levshina), but since that data comes with matching timelines the aligning task is very different.

However, there are also resources which are parallel in less obvious ways. There are historical events which have been covered in every newspaper in the Soviet Union and beyond. The data does not form the parallel sentences as such, but still those texts are thematically linked in very intriguing ways. As mentioned, applying tools such as named entity recognition into this kind of data could be very interesting for both language documentation and language technology, as well as to ordinary users and wider research in related fields such as anthropology.

Even more abstract parallels can be drawn between spoken and written resources, since there are many stories and narratives which have been recorded in different retellings and also published multiple times. A good example of this is the folktale Zarńia Bözha Kań, which also connects to the Russian folktale Maša i Medved. It is published in two different variants which are both stored in Komi Nebögain collection in Syktyvkar, and also has been recorded by Erkki Itkonen, as told by Vasiliy Lytkin in 1957 (Kotus recording id: 1323_2az). It is still unknown how this data can exactly be used, but as far as I see there are lots of open roads which can be explored.

Closing words

The usability of different workflows related to OCR and reuse of the materials is closely connected to copyright. From my perspective the best and clearest solution is to have as much data as possible in Public Domain. Even the use of different academic licenses are very problematic, since they do not correspond cleanly with commonly used open licenses, and it is unclear if they allow, for example, further derivations. As far as I can see, the majority of more interesting use would include creating derivational datasets of some sort. Only open enough license means that the data can be used in any possible way we can or will be able to imagine in the future. Naturally the existence of data in public domain is often related to the age of the data. This is why many large text corpora exist now for pre-1920's data, and as the years go by this year will gradually keep advancing. That said, it creates a situation where almost none of the presenters in this conference will live to see the currently published texts to enter Public Domain. There are exceptions such as the texts released into Public Domain by Ivan Belykh.

It is easy to talk about the differences in the data different projects deal with and produce, but there is a lot that is essentially same in all OCR workflows. These similarities mean that most of the practices and solutions which work in one place will also work elsewhere.

There is some kind of a boundary between modern printed resources and early modern manuscripts and prints, and lots of tools are customized to work with one or another. I also outlined above that the needs for presentation are clearly different with the handwritten manuscripts and printed products. This said, the boundary is not that clearly cut. The researchers are also today producing hand written fieldnotes, handling of which does not essentially differ what what we have to do for the notes of the figures at the Castrén's time. This is also obvious in the work we are doing in INEL. Kuzmina's and Donner's fieldnotes share between one another more than they differ, although there is more than a century differentiating them. Similarly early phonogram recordings can be dealt conceptually in rather similar way as the modern multimedia recordings.

References

Pääkkönen, Tuula, Jukka Kervinen, Asko Nivala, Kimmo Kettunen, and Eetu Mäkelä. 2016. "Exporting Finnish Digitized Historical Newspaper Contents for Offline Use." *D-Lib Magazine* 22 (7/8). http://www.dlib.org/dlib/july16/paakkonen/07paakkonen.html.

Tafti, Ahmad P., Ahmadreza Baghaie, Mehdi Assefi, Hamid R. Arabnia, Zeyun Yu, and Peggy Peissig. 2016. "OCR as a Service: An Experimental Evaluation of Google Docs Ocr, Tesseract, Abbyy Finereader, and Transym." In *Advances in Visual Computing: 12th International Symposium, Isvc 2016, Las Vegas, Nv, Usa, December 12-14, 2016, Proceedings, Part I*, edited by George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Fatih Porikli, Sandra Skaff, Alireza Entezari, et al., 735–46. Springer International Publishing. 10.1007/978-3-319-50835-1_66.

Wagner-Nagy, Beata, Hanna Hedeland, Timm Lehmberg, and Michael Rießler. 2015. "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.* Hamburg: Projekt Geisteswissenschaftliche Infrastruktur für Nachhaltigkeit (gwin). https://lecture2go.uni-hamburg.de/l2go/-/get/v/18306.