Challenges in OCR today: experiences from INEL

Niko Partanen

31 January 2017

Introduction

This paper discusses some experiences and workflows from INEL project (https://inel.corpora.uni-hamburg.de) (*"Grammatiken, Korpora und Sprachtechnologie für Indigene nordeurasische Sprachen"*). 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) (http://www.dlib.org/dlib/july16/paakkonen/07paakkonen.html). 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 human subjects. However, in part this must be connected to lack of knowledge and good practices with the licensing and data sharing. This is very unfortunate, since incorrectly structured and licensed materials will take considerable future effort and negotiation before they can be taken 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 text data. 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 gaps 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 (http://eap.bl.uk) ran by the British Library, have focused mainly to audio and photograph mediums. Others, namely Fenno-Ugrica (https://fennougrica.kansalliskirjasto.fi/) and Komi Kyv (http://komikyv.ru) website maintained in Syktyvkar, are good examples of 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 languages 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 later. Alongside Komi-Permyak, also Tundra Nenets likely falls into category where the number of available publications is particularly high. Indeed, with smaller languages it is likely that the recorded and transcribed materials exceed in size the printed ones, which can be used as one benchmark about the general state of resources in given language. The texts we work with in INEL 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 good solution to 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, especially 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 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(https://transkribus.eu), 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/software) has their own. PoCoTo (http://thorstenv.github.io/PoCoTo) has been developed in CIS Munich within a larger EU project, and uniquely from the others, it seems to have features related to batch correction of systematic mistakes. Also from this point of view one has to recognize character or word location on the page as an essential part of the OCR result, since it is impossible to imagine a post-correction workflow which would function without this information.

Compared to this, the initial training of an OCR model is very well suited to command line environment, especially for the transparency and reproducibility of the training process.

Current solutions

The recurring problems

Although there are several tools, many of them still tend to lack some features, the most critical of which are:

- Lack of search and replace with regular expressions (Mancinelli 2016, 259)
- Lack of good XML export

The first issue connects to automatization of the proofreading process, since some mistakes are always systematic and could be easily corrected before manual phase. The export format causes a more serious issue, as it connects to post processing and reuse. Following attributes would ideally be present in the output format:

- Recognized word (or character)
- Coordinates on page
- Recognition confidence
- Some style information
- Information about the used OCR system
  - Version
  - System settings
  - OCR model used (with version)
  - Information about page processing, if that was performed

As a format some kind of an XML seems to be the current standard, although expressing same information in other formats such as JSON would also be unproblematic. However, often offered formats such as plain text file, Word document or PDF are clearly not satisfactory. The National Library of Finland has also developed their own export format which contains ALTO XML, metadata and the text in plain text, and which was created to the need of users who benefit from access to local copies (Pääkkönen et al. 2016) (http://www.dlib.org/dlib/july16/paakkonen/07paakkonen.html). It can be expected that new solutions to distribute this kind of data will emerge, since the users certainly will have need to download the data locally and to update it with newest resources when they are added.

At the moment the OCR market is dominated by ABBYY FineReader (https://www.abbyy.com/finereader) and Tesseract (https://github.com/tesseract-ocr/tesseract) (Smith 2007). The first is used with good success in many large projects, for example in Fenno-Ugrica collection of Finland's National Library, but it is entirely commercial software which makes it very problematic from perspectives of open science and software. Tesseract, and its derivations such as OCRicola, are entirely open source and highly trainable, but they offer just the very core functionality of OCR recognition and thereby have to be used in integration with different tools, and at least in INEL we haven't found any good combination of tools which could offer a highly reliable and reproduceable workflow.

The main problem with ABBYY based solutions is that the two versions of this program, Desktop and Engine, are very strongly apart from one another and it is not possible to combine their functionalities. The main functionality which Abbyy Desktop lacks is XML export. This means there is no export format which would directly store the page position information. However, in the Desktop version it is possible to train new language models. I will cover this more in detail below, but within this the problem is that the models trained Abbyy Desktop cannot be used in Abbyy Engine, which would have needed XML export. Also it is certain that many users would not want to rely on desktop environment, but would like to automatize their work in different ways, which in principle could be done with Abbyy Engine. However, with rare scripts and endangered languages the current language models are highly insufficient and some additional training is necessary in order to achieve satisfactory output. It would already be a half way solution to be able to use the models trained in Abbyy Desktop in Abbyy Engine.

It is a good question why Abbyy Desktop doesn't have XML export. For a long time I was suspecting this was because of capitalistic business interests: there are some reasons why it is more lucrative not to offer XML export for a casual user. Maybe it would allow them to do too much, instead of relying on their more expensive options. However, after using Abbyy for a longer while I've started to assume there is a softer reson. Maybe even Abbyy doesn't know how to store user edited position information? When the user edits the text in Abbyy, there are moments where the word entirely loses the highlighting, as if the software would not have information about the position any longer. There are similar problems with the Revizor editor developed by National Library of Finland in Fenno-Ugrica project. In Revizor the information is usually kept, but especially at the line boundaries it seems to lose the knowledge about which words were originally on which line. This all signals that is must be extremely difficult to store word position information in user edited text. The assumption is very logical in the sense that the user can easily copy and paste text around, delete some words and retype them, and somehow the software would still need to keep track of positions. If the task is too difficult for the industry leader, we may in a deeper trouble than we thought, as it is unlikely smaller open source projects could reinvent this task. On the other hand machine learning and neural networks based tools are advancing now so fast that many tasks we currently think are challenging may find surprising solutions in the coming years. I remain optimistic that in the longer run the open source community will emerge as the winner. One demand for this is certainly to stop building a new software in every project and cooperate more across different involved parties.

In INEL we are currently working through Abbyy Desktop plain text file and HTML exports, which both keep well the paragraph structure and, with some manual maintenance, also page numbering. With the texts we have worked on it has been possible to distinguish page numbers as more or less sequantially growing numbers as only items on the line, and the chapter numbers have been on the same line with title or the first line of the text, in any case, not alone. However, this kind of definitions are not very generic and also needs continuous manual adjustation. The data retrieved this way is aligned with Russian sentences using hunalign (https://github.com/danielvarga/hunalign), again with some manual supervision, and converted

into a Toolbox file with a simple script which transforms the file structure to fit Toolbox demands. This is eventually imported to FLEx, line, chapter and page information being stored on note tier. This workflow disregards the page position information, which is suboptimal in every sense, but for now there have not been better alternatives. The solution selected in INEL works in careful digitalization of individual texts one by one, but would not be satisfactory in large scale digitalization projects.

Basics of OCR model training

OCR model training can be on different levels, and Abbyy and Tesseract in our use employ opposite strategies. This is a matter of level on which the recognition is being performed.

```
character < character combination < wordform <  morphological model
```

The Abbyy, as we have used it with minority languages, has been used by us on character and character combination level. Since the wordform information is not used at all, there is no interference from the Russian word frequency and character combination information which the software could use with Russian text. New combinations can also be trained, which can be especially helpful when teaching Abbyy to recognize rare and new characters which are not in Russian orthography. The next step would be to feed from Dolgan corpus to Dolgan Abbyy model, but one factor that makes this complicated is that different publications use different enough scripts and phological interpretations for each language that the modeling often has to be done individually for each publication. One benefit from this relatively shallow model is that it does not get mixed up by dialectal variation in the wordforms, as it doesn't look beyond individual characters. However, if the model has exact information about all possible wordforms, the recognition quality can possibly be improved even further. However, according to Silfverberg and Rueter the word lists may still be superior to morphological analysators (2014). It must also be taken into account that more sophisticated model may also struggle more when the text it encounters is less standard. In later step the language data will be included better to the model, since in INEL all transcriptions are eventually converted into phoneme level transcription system which is used consistently within the project, and in some cases conversion to the variant used in publications should also be possible, as long as the phonemic level is same.

Training an Tesseract/OCRicola model

I did rather comprehensive testing of OCRicola and Tesseract in fall 2015, and especially OCRicola offers the opposite approach to what we are now doing with Abbyy. Tesseract model training is based to the idea that an example images are generated from an example text and a font. OCRicola is a software built on basis of Tesseract and has most of the same functionality. The main difference between the tools is that OCRicola is able to use finite state transducer (HFST) to generate the word forms used in recognition. In theory this approach is superior to corpus based input, since the morphological analysator is able to generate also the forms which are never present in the corpus but may occur in a new text. One issue related to this is that OCRicola was split from and older Tesseract version, and currently we are already in Tesseract 4.00. As these are both open source projects, this is not necessarily a problem. However, for example, the newest Tesseract already includes a neural network subsystem as a new type of textline recognizer. This suggests that over time there will be very interesting new features in Tesseract. One could think that the most exciting new features would be usable only with larger languages, but as the Komi Zyrian corpus is also now over 30 million words, I see no problems in adapting almost any approach which works with larger languages to Komi as well.

Challenges with Tesseract/OCRicola model are related especially to the fonts and availability of large enough training corpus. With OCRicola the question arises also about the existence of available Finite State Transducer, but at least within the Giellatekno infrastructure a large portion of Uralic languages are already covered to different degrees.

Tesseract is easy to use and install, but to get all dependencies installed correctly for training new models can be very complex and differs between versions. Probably this is not an unusual problem, and the installation

is indeed possible and works, but this cannot be under any conditions described as straightforward or unproblematic.

Training an Abbyy model

In contrary to Tesseract/OCRicola model, Abbyy takes care most of the things by itself without user interference. This can be seen as both an up or downside, since it makes the model development entirely non-transparent. This allows using other Cyrillic model, such as Russian, as a backbone and build upon that by adding the special characters needed. The new characters can be added and those which are not needed removed, and Abbyy has well functioning tool to manually assign different character forms to the model as representative of specific character. This makes Abbyy model training very generic and allows even arbitrary matching between character and recognized forms, which is a great benefit when compared to Tesseract/OCRicola, which demands that the trained character is also present in the font.

It is also possible to add dictionary data to the Abbyy model. This is something we have not tested, but which larger languages are clearly having in their models. Indeed, recognition of languages such as English, Russian and German is almost flawless. In the same vein it must be stated that also with new languages Abbyy often reaches very good results with relatively small investment to the training. The only concrete problem with Abbyy is the lack of transparency behind the model training, lack of interoperability between different versions and some issues in missing export formats.

One additional issue is that only Abbyy Engine is able to produce ALTO XML, which many libraries are using as their standard format for all data. This is an open format, but this creates a setting where it is not possible to store the manually corrected versions in the preferred format. Due to this the Fenno-Ugrica collection also contains better versions in plain text files and worse recognition in XML files.

Approaches to the digitalization

I have often encountered the thought that the page position information would not be needed. However, there are several situations where this information is very critical. For example, different page elements such as page numbers are very easy to recognize from the layout when their position is known, but tend to get very messily mixed with rest of the content when the position information is lost. The same goes with paragraph information, since a paragraph is basically a structure in text which is closely related to the organization of line indent, which, when known, makes paragraph detection rather reliable. Detecting it from unstructured lines is, on the other hand, rather hopeless.

It can be argued that there are special cases where this information can be disregarded. For example, when working with texts like Bible or Koran the original verse structure is usually enough to distinguish different elements, and in many instances the sameness of the text in each version is very highly enforced. Some texts resulting from OCR, for example elicitated sentences or structured questionnaires, follow essentially the same model where each utterance can be identified as parts of these closed sets. Some data encountered in INEL is also like this, for example the translated sentence lists collected from many different languages in Siberia. This these sentences the number and the information that they belong to this specific set is enough to identify them reliably.

Thereby OCR can be seen as a mean to get the text, but not so much as effort to digitalize the entire individual book or manuscript. Only later post processing turns the text into research data. More rare and more subjective to reading and interpretation the text is, more valuable it is to be able to reproduce the original text or to easily match the recognized words to positions on the page. In this sense OCR is ultimately just an attempt to claim that on these pixels, or on these coordinates on the original page, we have representations of these specific characters. Especially with old typefaces or handwritten manuscripts the matter of interpretation tends to become more significant.

There are other good discussions about the steps in OCR process, for example Holley (2009) goes it through step by step. Although that description is some years old, the principle has not changed, and the table

1 in her paper presents well all needed phases (http://www.dlib.org/dlib/march09/holley/03holley.html). What I would mainly want to emphasize here, is that it is somewhat questionable how well the current tools actually store information about what exactly has been done in which point. The measure of this would be whether we are able to associate an arbitrary token in the final recognized text with the matching coordinates in the original manuscript, microfilm or whatever the source has been. At least in regard to INEL workflows the answer is unfortunately no.

There are some experiences that the OCR resulting text can be corrected to some degree already with simple ngram-based post-processing (Arkhangelskiy and Medvedeva 2016). This is certainly true, especially with data which is relatively well recognized, although the worst recognized parts should probably be approached with redoing OCR (Kettunen and Pääkkönen 2016). I would still be cautious about not having the need for manual post-correction, which although impractical for millions of titles, can still be relatively cost-effective for hundreds or thousands of proofread books in a year. Needless to say, already this can produce a very large corpus, as the example of our colleagues in Syktyvkar shows.

Within INEL, the OCR is not a direct goal in itself, but the text extracted from the documents will be associated with its original recording. There are also cases where the original recording is lost or was never made, which is the case with older transcriptions. In the latter case the OCR'd document is the actual primary data source, or the closest representation of that which we can have. Thereby there has to be relatively close connection between the original text document and the later XML file. The currently used XML format in INEL is the one produced by EXMARaLDA. The tools within EXMARaLDA are mainly meant for work done with spoken language data, but we have had good experiences in using it also with the written data. The use of tools like these is inevitable since we cannot store in plain text files all the further annotations we want to provide. In Hamburger Zentrum für Sprachkorpora we have already a long practice to publish online different visualizations for spoken data (see Demo corpus on the HZSK website (https://corpora.uni-hamburg.de/drupal/de/islandora/object/spoken-corpus:demo)). During 2016 we have also tested different visualizations for interlinearized text which contains no multimedia, and also for this there are clearly many good options.

Publishing the texts is naturally much more simple when there are no annotations. One of them is simply rendering the text as plain HTML. For a normally structured book this already makes the reading experience very enjoyable. On the other hand if it is possible to render the text into HTML, it is also possible to produce other formats such as EPUB which are very good to read on mobile devices.

There are also some types of documents which would benefit from more complex rendering online. Especially this is the case with different manuscripts where even the reading of different characters may be debatable. Although I don't discuss the text recognition of hand-written documents here in depht, it should be pointed out that this is the type of document where some new facsimilé type of online publishing options would be the most needed. For example, the manuscripts of M.A. Castrén have recently been edited in ISO-TEI facsimilé in Helsinki, and the data of many other reseachers is still unprocessed in the libraries and museums, among those also researchers of Komi such as Sjögren and Sirelius. I'm certain also wider audience would appreciate the possibility to examine these materials easily online. What I would like to see the most is a set of tools which allow very easy and shallow rendering of XML and middle quality images, instead of building the solution on top of heavy servers and byzantine architectures which demand complex customization to fit the needs of specific projects. As I've discussed here, those needs may in reality be more similar than we think. The XML files could be even stored in public repositories such as GitHub, which would bring the work more towards actually open and collaborative science.

In an ideal world the OCR model development could be done with Open Source tools for languages which have the following resources:

- Large corpus available with a permissive license
- A good collection of fonts which match closely the ones used in publications
- Frequency lists from possibly even a larger corpus
- Consistent spelling which doesn't vary across publications

It is noteworthy that for the large languages such as English or Finnish all these resources and conditions exist. However, for many of the smaller languages these form a large issue. Thereby it is also advisable

not to publish corpora with licenses which contain clauses such as No-Derivations, since an OCR model is a derivation of one sort. Although it may be largely untested what exactly constitutes as a derivation, I would not start to use such a corpora in this kind of work.

Challenges for language technology

OCR'd texts pose many forthcoming tasks for natural language processing. First of all, there is the need to transform highly idiosynchronic writing systems into same ideally phoneme level representation. Moreover, OCR'd text is often not split neatly to individual words, but somewhat specific form of tokenization has to be employed. In the same vein one could also ask for reliable detokenization, since the result of OCR is often not really the sentences, but not really the word tokens either. The resulting text is in many ways complex, and the ready tools may not be adjusted to work exactly with that.

The written text often needs some kind of a morphological analysis in order to be more usable for linguistic analysis. Thereby applying different tools to this end is highly desiderable. For example, with digitalized Komi-Zyrian and Saami materials GiellaTekno tools have already proven very useful, and maybe something similar could be applied later also to INEL corpora. It can also be mentioned that many texts are attractive targets for named entity recognition (NER) since they often cover repeatedly same prominent historical events in different languages and repeat same historical figures.

One particularly relevant tiny task I can foresee is the automatic matching of recognized text with the original pages. This can be done by comparing automatically the proofread text masses with the XML output with word coordinate information. The two texts are naturally different, as one variant is not proofread and contains many errors which never have been present in the other version. However, I believe this is a minor obstacle, and detecting the most similar pages (especially knowing that the following page normally continues the text) should be very doable in immediate future. I have not found any existing solutions to this, which again may reflect the immaturity in the use of OCR'd research data in the fields where high quality text output is a necessary or high preference.

Contribution to language documentation

In corpus linguistics and language technology the texts are usually treated as the target of investigation in themselves. When applied to the endangered languages, there is often need to personalize them further and add into metadata information about the writers and translators, among other relevant factors. Instead of variables such as publishing time and place there is obvious need to know writer's birth time and birth place. Already those tell quite a bit about the possible native dialect of the writer, which again can be one way to explain idiosynhcronic features of individual texts. Often this kind of information is stored already in public repositories such as Wikipedia, so it is one additional question how to employ these connections the most effective way.

In many endangered language communities the pool of active and prominent speakers is small. It is highly likely that the same individuals have been writing to local newspapers and even published longer pieces of prose, and have also been speaking in public events which have been recorded. Or they have ended up to be recorded by different linguists of ethnographers, resulting in archive items which can be relatively easily associated with each other. Similarly in case the writers have deceased it can be expected that many communities recognize and remember them.

One also has to take into account that as the number of available resources grow, it is often not realistic that the research is conducted with one specific corpus, but there are many possibilities nowadays to compile a new dataset from many different corpora. When the data is archived with proper metadata it should even be possible to harvest metadata from distinct sources and select carefully those recordings and transcriptions which contain specific criteria there are for research question at hand. What I try to explain is that even when the corpus would contain a large number of texts which someone would not find interesting or relevant, nobody is forced to use the complete corpus. Thereby building besides the corpus a large collection of OCR documents in different quality does no harm, but is one more asset that can be used if the need arises.

Last it could also be speculated that with many languages the existing language documentation resources are the largest available body of texts in that language. In these cases one could also imagine scenario where the language documentation data could be directly used to create an OCR model. This is somewhat counterintuitive, but text is text, and if the match between phoneme representation in written and spoken varieties is close enough, which it often is with languages with new orthographies, the changes for this to succeed are high.

There are also additional research possibilities with the parallel texts. Within Fenno-Ugrica project it has clearly been one of the corner ideas to collect resources which exist in several languages spoken in Russia. Thereby there are now publications such as Four Battles, which are translated to nine different Uralic languages. A cursory Google search reveals that translations must exist also in different Turkic languages. When the books are compared the texts appear immediately useful, since they share exactly the same structure, often almost on sentence level with only small deviations. Extracting the parallel sentences can be done with customary command line tools (Perl, sed) and, for example, hunalign, which automatically detects matching sentences and outputs a format that can be used in further processing. However, aligning the sentences is currently somewhat time consuming and more elaborate workflows are desperately needed. That said, matching the sentences semi-automatically is rewarding and fast enough that doing it manually for individual books is entirely feasible. Recently also a corpus of movie subtitles has been compiled (ParTy 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

Arkhangelskiy, Timofey, and Maria Medvedeva. 2016. "Developing Morphologically Annotated Corpora for Minority Languages of Russia." In *Proceedings of Corpus Linguistics Fest 2016. Bloomington, in, Usa, June 6–10, 2016.*, edited by Sandra Kübler and Markus Dickinson, 1–6. CEUR Workshop Proceedings 1607. Bloomington: Indiana University. http://ceur-ws.org/Vol-1607/arckhangelskiy.pdf.

Holley, Rose. 2009. "How Good Can It Get? Analysing and Improving Ocr Accuracy in Large Scale Historic Newspaper Digitisation Programs." *D-Lib Magazine* 15 (3/4). http://www.dlib.org/dlib/march09/holley/03holley.html.

Kettunen, Kimmo, and Tuula Pääkkönen. 2016. "Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled Ocr Data with Basic Language Technology Tools and Means." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (Lrec 2016)*, edited by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, et al. European Language Resources Association (ELRA).

Mancinelli, Tiziana. 2016. "Early Printed Edition and Ocr Techniques: What Is the State-of-Art? Strategies to Be Developed from the Working-Progress Mambrino Project Work." *Historias Findigas* 4: 255–60.

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.

Silfverberg, Miikka, and Jack Rueter. 2014. "Can Morphological Analyzers Improve the Quality of Optical Character Recognition?" In *Proceedings of 1st International Workshop in Computational Linguistics for Uralic Languages*. http://dx.doi.org/10.7557/5.3467.

Smith, Ray. 2007. "An Overview of the Tesseract Ocr Engine." In *Proceedings of the Ninth International Conference on Document Analysis and Recognition (Icdar)*, 629–33. The IEEE Computer Society Conference Publishing Services. http://www.australianscience.com.au/research/google/33418.pdf.

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.