Final report project Full Automatic Archival Access (FAAA)

Ministerie van Volksgezondheid, Welzijn en Sport

VSBfonds, iedereen doet mee

vfonds investeert in VREDE

BRAIN Branchevereniging Archiefinstellingen Nederland

Amsterdam 13 October 2016

Contents

The project Full Automatic Archival Access is part of the Archief2020-programme (www.archief2020.nl). It was funded by Archief2020, BRAIN, VSBFonds, VFonds and the Ministry of Health, Welfare and Sport (VWS).

Summary

Aim of the project Full Automatic Archival Access was to find out more about the possibilities of currently available proven technology to create useable machine-readable text out of analogue archival typed or hybrid documents. The project partners – Netwerk Oorlogsbronnen (Network War Collections), Nationaal Archief (National Archives of the Netherlands), Impact Centre of Competence and Centre of Language and Speech Technology – compiled a small non-representative sample set from the Centraal Archief Bijzondere Rechtspleging (National Archives). The tests focused on the best settings for scanning, pre-processing for Optical Character Recognition (OCR), transcription (OCR) and post-correction (in particular named entity recognition), using state-of-the-art software (e.g. Abbyy Finereader) and common practice tools (e.g. TICCL and FROG).

For the test set the best OCR-scores were produced when:
- the documents were captured with a black background
- the documents were sequentially deskewed and the border was removed before OCR'ing
- the documents were processed by Abbyy Finereader 11 SDK

With these 'ideal' settings a weighed word accuracy rate was measured of 81%, meaning that roughly four out of five words (out of a total amount of words of about 30,000) were correctly transcribed by the software.

Some additional tests with the post-correction tools TICCL and FROG gave a glimpse of some promising possibilities to correct incorrectly transcribed names of persons, places, organizations, events and domain specific WW2-related nouns ('deportatie', 'interneering' etc). Feeding the software with existing lists and ontologies seems to be a valid strategy for further improvement. Apart from that further fine-tuning of existing tools is likely to increase the quality of the post correction and lower the error rate.

Based on the results of this small project full automatic transcription of typed or hybrid documents with state-of-the-art tools seems to produce machine-readable text that is good enough– with a relatively small error margin – for providing digital access on document level. Improving this machine-readable text with post correction tools seems very promising. This technology requires further development and is currently still quite experimental. However, if it is possible to reach a word accuracy rate of 81% with full-automatic means, you can already provide a high level of access without too much effort and costs. It is important to stress that it does not work for all kinds of typed or hybrid documents: especially documents with straight lines, portrait perspective, strong contrast between text and background and a stable ink density are most suited for OCR'ing.

The tests were carried out between December 2015 and October 2016. The project was funded by: Archief2020, BRAIN, Ministry of Health, Welfare and Sport (VWS), VFonds and VSBFonds.

1. Background

According to a survey held in 2014 by the EU-funded project ENUMERATE only 8% of the Dutch archives have been digitized.[1] The larger part of these digital archival collections have not been made searchable full text on document level. The main reason for this is that archival documents – often fragile papers with complicated lay-outs, bad printing quality, in many cases handwritten text that is hard to read even for humans – are complicated for text recognition software to recognize. In case of printed works (books, newspapers, magazines, etc.), due to substantial improvements of the quality of the software the last few years, OCR'ing of the full contents has become common practice (Delpher, Google Books). For archival materials, OCR'ing or HTR'ing (Handwritten Text Recognition) have been applied on an experimental basis, but little is known about the current potential of state-of-the-art text recognition software.

In the field of Digital Humanities – "useable" – big data corpora are scarce. Many scientifically eminent heritage collections remain inaccessible for new digital research methodologies. At the same time, often in experimental settings, tools are being developed that could improve access to collections considerably. For instance, named entity recognition (NER)-tools can be applied to automatically produce indexes for person names, organizations and geographical locations. Post-correction tools can help to improve corrupt OCR by comparing significant words with historical lexicons with persons names (e.g. from population census), geographical names (e.g. historical Geocoder), names of organizations (e.g. lists collected by Netwerk Oorlogsbronnen) and domain-specific resources (e.g. Dutch Thesaurus WW2) and spelling variants.

In 2015, at the initiative of Netwerk Oorlogsbronnen[2] – a national cooperation of holders of Second World War-collections in the Netherlands focusing on improving digital access to the approximately 400 collections in the country – the project Full Automatic Archival Access emerged. A joint project proposal to do a small pilot project together with the National Archives, the IMPACT Centre of Competence and the Centre for Language and Speech Technology (see appendix) received funding in November 2015 from the innovation programme Archief2020[3] and the branch organization of archives BRAIN.[4] The project ran from December 2015 to October 2016.

2. Scope and aim

The main aim of the project Full Automatic Archival Access was to find out more about the possibilities of currently available proven technology to create useable machine-readable text out of analogue typed or hybrid[5] archival documents. Since the primary focus of Netwerk Oorlogsbronnen is on opening up archival collections that typically consist of typed or hybrid (mixed) documents, it was decided to exclude handwritten documents. In order to judge its potential objectively, standardized quality measurements on specific points in the workflow were made. Because of the limited resources of the project we focussed our research on a relatively small sample that would at least provide us with indicative figures on the quality.

[1] www.den.nl/art/uploads/files/Enumerate-core-survey-NL2013-2014.pdf.
[2] www.oorlogsbronnen.nl/programmaplan2015.
[3] www.archief2020.nl.
[4] www.archiefbrain.nl.
[5] Partly typed, party handwritten.

A small test set was compiled from the Centraal Archief Bijzondere Rechtspleging (CABR), held by the National Archives. The CABR, consisting of approximately 4.5 kilometres, is the most frequently consulted WW2-archives in the Netherlands. It holds a wide range of various documents (eye witness accounts, legal documents, correspondence, verdicts) in relation to some 300,000 persons who were under suspicion of collaboration with the German occupier. Due to privacy legislation the archives will not be fully open until presumably 2025. Currently the CABR is only accessible by employees of the National Archives on the name of the suspect and his date of birth. The documents of the CABR have not been digitized. Especially the judicial documents and eye witness accounts contain a wealth of information on not only the perpetrators, but also the victims. More in general it is a goldmine for any researcher looking for information on specific events, locations or persons in relation to the Second World War in the Netherlands.

3. Methodology and results

In a digitization workflow that starts with a paper original and ends with a machine-readable text file the next phases can be distinguished:

A. Selection and preparation of the source documents
B. Scanning
C. Pre-processing for OCR
D. Transcription (OCR)
E. Post-correction

A. Selection and preparation of the source documents

The CABR consists of 4.5 kilometres of archival documents. Within the scope of this project it was impossible to perform a collection survey and put together a statistically valid sample set representative for the collection as a whole. A total number of 99 documents (typed and hybrid) were selected from two different inventory numbers (CABR 2.09.09 548 and 2.09.09 542). According to the collection specialists of the National Archives the sample set contains documents that occur frequently in the whole of the CABR.

Documentation

- Selection invnrs CABR pilot (PDF)

B. Scanning

Methodology

The selected files were scanned

- with a black background in context (stack of papers)
- with a black background
- with a white background

- with a light grey background
- with a middle grey background

They were processed at a resolution of 300 ppi in colour, in TIFF 6.0, according to standard specifications of the National Archives. There was no post-processing applied. The converted JPEG-files have a baseline compression 1: 10.

IMPACT-Centre of Competence measured the effect of different backgrounds on the quality of the OCR.

Conclusions

Images scanned with a black background produced the best OCR-results.

Documentation

- Comparison of different settings for the digitisation of CABR by IMPACT (PDF)
- Offerteaanvraag digitaliseren CABR door Nationaal Archief (PDF)
- Testsetspecs Bijlage J Eisen (PDF) (zie de eisen voor het kavel: archieven, standaard)

C. Pre-processing for OCR

Methodology

Tests have been performed with 89 documents[6] in Abbyy FRE 11 and Abbyy FRE 12 Pro. Some border removal (removal of non-textual regions surrounding an image) and deskewing (straightening of scan to get horizontal text lines) tools have been applied, as well as binarisation (transforming colour into black-and-white e.g. to reduce the effect of bleed-through) of the colour tiff files. In order to be able to measure the output of all 89 files have been manually transcribed and used as ground truth documents.

Conclusions

- The best OCR results were measured in the application of the tools in the sequence: deskewing – border removal – OCR. Binarisation – often applied to improve OCR of printed matter - produced higher error rates because the ink density of the documents is less homogenous than in printed documents. Also, feeding the Abbyy Finereader 11 SDK with gazetteers provided by the Instituut voor Nederlandse Lexicografie and German/Dutch names in the Geonames-database did not lead to better results.
- Abbyy Finereader FRE 11 SDK produced better overall OCR-scores than Abbyy Finereader 12 Pro.

Documentation

- Comparison of different settings for the digitisation of CABR by IMPACT (PDF)

[6] Ten documents less than were digitized because there were some identical documents.

D. Transcription (OCR)

Methodology

89 files were OCR'ed with Abbyy Finereader FRE 11 SDK and Abbyy Finereader FRE 12 Pro. Ground-truth files were manually created with the IMPACT OCREvalution-tool[7] to compare the results with gold standard-documents.

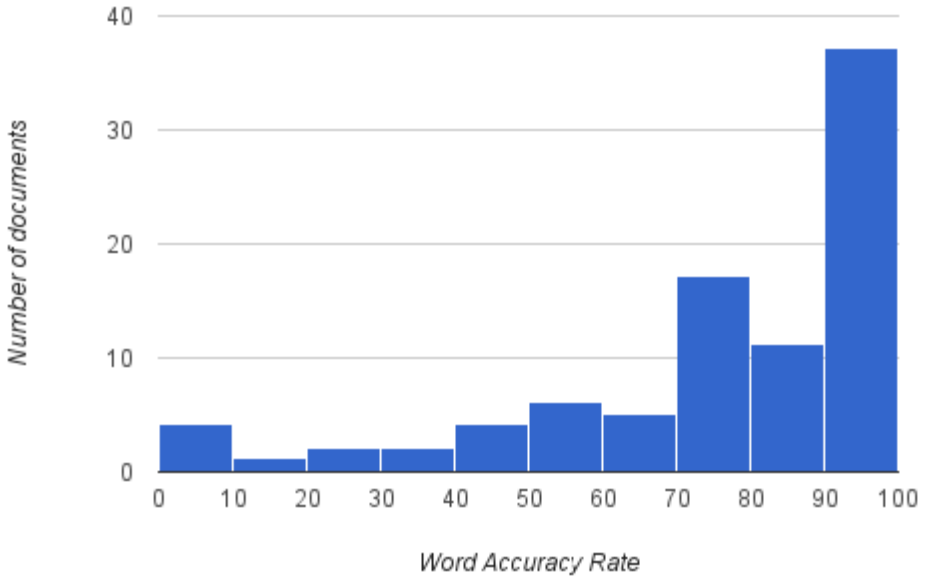

Figure: distribution of word accuracies (X = Word Accuracy Rate Y = number of documents.

Conclusions

- The transcription of most documents in the test set reaches an accuracy rate over 80%. There is a small fraction of documents of which the accuracy rate drops below 10% (see figure). The weighted average word accuracy is 81,06% (order independent)[8] or 75,95% (order dependent).
- Best scores are with typed documents with:
  - text that has a homogenous ink density
  - a portrait layout (no tables or complex forms)
  - a solid, undamaged paper base (no bleed through)
  - a strong contrast between well-defined text areas and a light background
- Abbyy FR11 SDK produced better OCR-results (in the ideal workflow) than Abbyy FR12 Pro.

[7] https://github.com/impactcentre/ocrevalUAtion

[8] Order independent aka 'bag of words' means: word appears on the page but not necessarily in the position indicated in the ground truth document.

Naam en voornamen: @ACHTERNAAM@,@VOORNAAM@ Geboortep!aats en-datum: **Zaandam**, **29 October 1897** Echtgenoote van / Beroep: voorheen **agent van Politie**, t Laatste woonptaats en adres: **Kanaalstraat 25 II Amsterdam** **Persoonsbewijs**-no.:z 2 01239 afgegeven te **Leeuwarden** Nationaiiteit (evt. vroeger) **Nederlander** die ervan **verdacht** wordt: **joodsche personen** in macht van den **vijand** te hebben gebracht,terwijl hij in dienst was van de **S.D.** Terzake gehoord, verkiaarde **verdachte** mij het votgende: dat hij **in dienst was getreden** van **Lippmann en Rosenthal** voor de inventarisatie van **joodsche goederen**,vervolgens overgegaan naar de **S.D.,afdeeling joodsche zaken** te **Amsterdam** Verdachte heb ik, optastvan den Chef **Opsporingsdienst D.P.M**. op **28 Mei 1945** **bewaring** , togesteid, in het **Huis van Bewaring I** te **Amsterdam** **P.O.D. Amsterdam.** **Mode! A**

Domain specific words (Second World War)
Persons
Organizations
Geographical location
Date

Figure: fragment from CABR-document with word accuracy rate (= % of correctly converted words) of 80,55%. The CABR archive is a rich information resource because of its numerous, often standardized references to time/date, person names, names of organizations and domain-specific terms and expressions.[9]

Documentation

- Comparison of different settings for the digitisation of CABR by IMPACT (PDF)
- Analyse CABR overall (xlsx)

[9] Due to privacy restrictions the names, dates and geographical locations in this example have been altered.

E. Post-correction (including named entity recognition)

Methodology

Gold Standard (GS)-versions for all 89 files were manually annotated. Focus was on identifying person names, names of organizations, geographical locations, other names (e.g. nationality or occupation) and dates or other time references. The GS-files were not aligned, so the exact position of the named entity has been left out of scope. Therefore the results presented are tentative and only indicative.

Three tools were used with named entity retrieval of the OCR'd selection of CABR-documents
- FOLIA: the OCR of the CABR documents was converted to FoLiA XML format (Format for Linguistic Annotation). FoLiA is a rich XML-based annotation format for the representation of language resources (including corpora) with linguistic annotations.
- TICCL: Text-Induced Corpus Clean-up or TICCL provides non-interactive spelling and OCR post-correction facilities.
- FROG: linguistically enriching the texts.

Several tests were done with different combinations and settings.

Conclusions

- In the majority of the cases, results in terms of numbers of named entities automatically retrieved, as compared to the numbers of manually annotated named entities as measured across the different classes, are good to very good. The relatively good results of the OCR-tests obviously also lead to better results with Named Entity Recognition tools. Fully automatic OCR post-processing with TICCL further enhances the overall quality of the text and the amount of valuable metadata in terms of names recognized and categorized according to the classes defined.

- For further improvement tools available for post-correction should be fine-tuned more to the specific characteristics of the CABR corpus. For instance, correction of the words on the basis of frequency (n-gram correction) would help to correct wrongly interpreted named entities. Specific training of FROG (developing it into a "prince") or TICCL is also likely to provide better scores. Further recommended follow-ups are to use other available tools for retrieving date and time (Heideltime), events and locations in a broader sense ("wood", "house", etc.) and monetary values.

Documentation

- Full Automatic Archival Access Named Entity Retrieval on CABR by CLST (PDF)

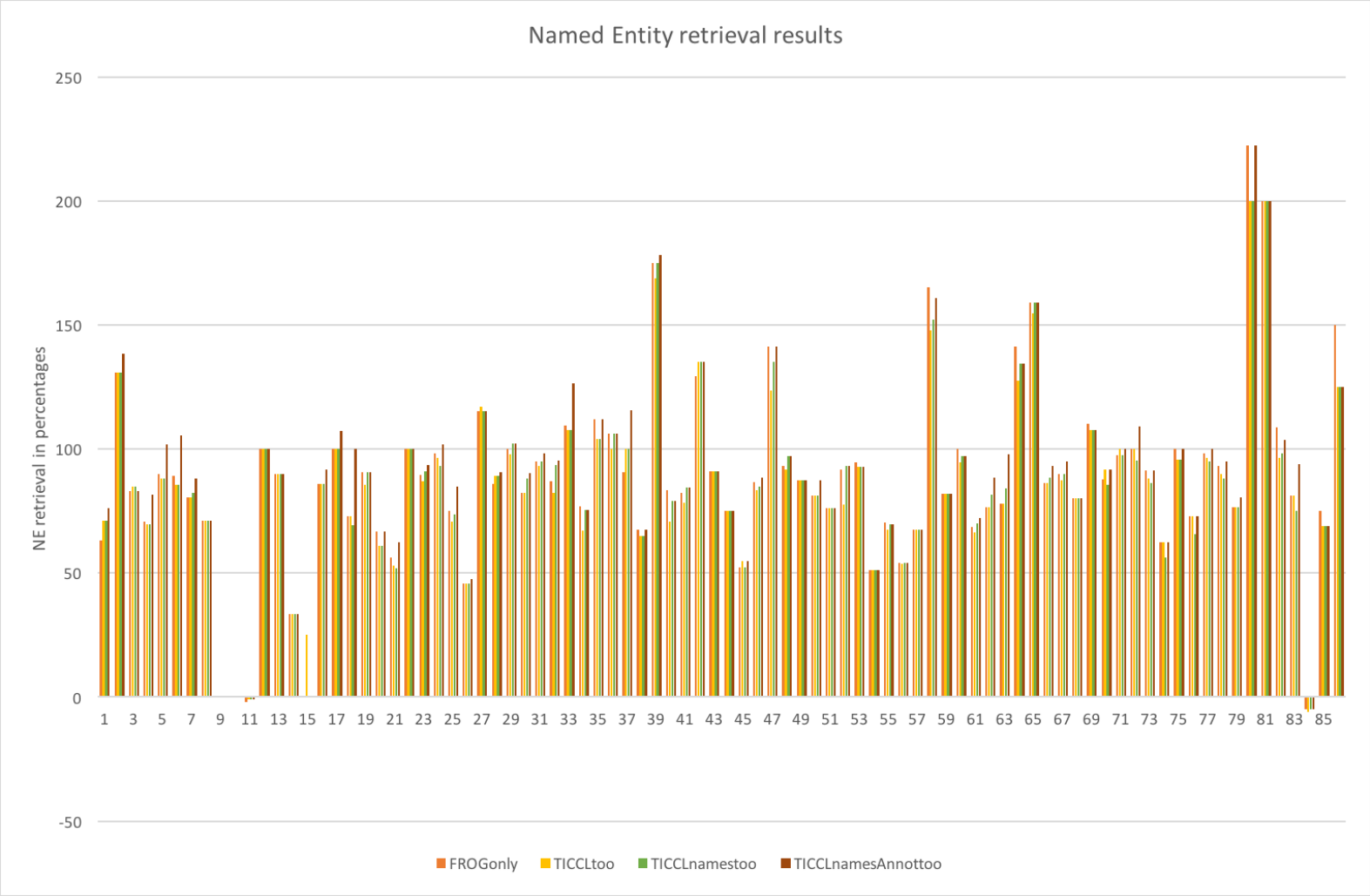

Figure: number of named entities retrieved in each experiment per class contrasted to the numbers annotated in the Gold Standard

Appendix A: project partners

Netwerk Oorlogsbronnen (www.oorlogsbronnen.nl)

- Edwin Klijn: project leader

IMPACT Centre of Competence (www.digitisation.eu)

- Rafael C. Carrasco: OCR-researcher
- Isabel Martinez: coordinator

Centre for Language and Speech Technology (www.ru.nl/clst)

- Martin Reynaert: post-correction researcher

Nationaal Archief – National Archives of the Netherlands (www.nationaalarchief.nl)

- Anne Gorter, Liesbeth Keijser: coordinator
- Joop Korswagen: digitization advisor

Appendix B: publications and conference

All documents mentioned are available at: www.oorlogsbronnen.nl/volauto

Test documents are protected by privacy laws. They have been deposited as research data set at the digital depot of the Nationaal Archief.

Publications

‘De experts aan het woord: wat kan de archiefsector met nieuwe digitale technologie?’, Archievenblad November 2016.

‘Googelen door archieven. Een revolutie in archieftoegang of sciencefiction?’, Archievenblad December 2016.

Conference

Archief2020 Conference ‘Googelen in Archieven’, 13 October 2016, Nationaal Archief: bit.ly/Verslag13okt16