Report on the comparison of different settings for the digitisation of the collections of Centraal Archief Bijzondere Rechtspleging (National Archives of the Netherlands)

Author: Impact Centre of Competence

Date of preparation: July 12, 2016

1 Motivation

This report is part of the national project *Volautomatische Archiefonsluiting*,[1] which aims to investigate the best procedure for the OCR (Optical Character Recognition, the process which automatically interprets textual content) of archival documents when they contain printed or hybrid text (hybrid text includes, for example, stamps or handwritten annotations).

The objective of this report is to determine the recommended workflow (a generic workflow is shown in Figure 1). and settings for best results when state-of-the-art technology for OCR is applied to the project materials. After the compilation of a test set, the results obtained under several configurations will be compared and analysed. The report will present standard indicators measuring the quality of the output under different settings as well as the conclusions drawn from their analysis.

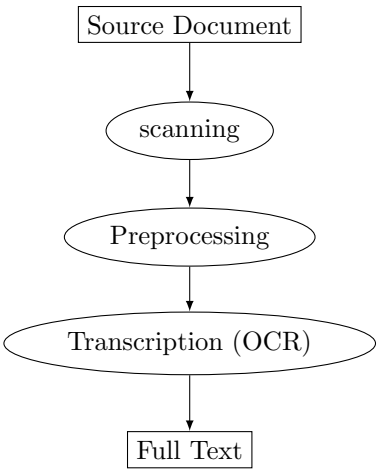

Figure 1: A generic workflow for the automatic transcription of text.

2 Source documents

A sample containing 89 documents (with group identifiers 542 and 548) were manually selected from CABR-archives of the National Archives by the National Archives team. These documents show complex and variable layouts and they contain mainly handwritten and typewritten scripts in Dutch with some German inserts (for detailed specifications, see Appendix A). These features pose a challenge to available layout analysis tools and to current OCR engines.

[1] Funded by Archief2020 and BRAIN, and led by NIOD (Institute for War, Holocaust and Genocide)

Every source document was scanned against different backgrounds: placed on a uniform surface (white, light grey, dark grey, or black) and also in the context it was found in the containing physical folder. The documents were scanned at a resolution of 300dpi and the images were generated both in TIFF and JPG format. The former Impact project (www.impact-project.eu) reported that the Abbyy FRE provides the best results when the input was scanned at this resolution (remarkably, higher resolutions did not improve the results).

3 Tools involved in the evaluation

According to the definition of this project, the Abbyy FineReader suite of tools (Abbyy FRE for short) will be used as state-of-the art OCR engine. The image processing tools were selected after the experience gained during the EU-funded project Succeed (http://succeed-project.eu). Succeed promoted the validation and take-up of tools and compiled information about digitisation software meeting a number of criteria:

1. It is freely available or, at least, a free trial version is distributed.
2. Either sufficient technical documentation is accessible or the code remains actively supported.
3. The software is endorsed by its usage in projects, existing benchmarks, or the information gathered from users.

The following tools meeting these requirement were selected and used in the tests described in this report:

- Image enhancement tools:
  - Image Magick Border Removal.[2]
  - NCSR Border removal[3]
  - Deskew Tools by Galfar's Lair (through Bitbucket).
  - Image Magick Deskewing
  - Abbyy FRE 10 Binarisation.
  - Image Magick Binarisation
- OCR Engines:
  - Abbyy FRE 11 SDK Version.

[2] http://www.digitisation.eu/training/succeed-training-materials/image-processing/image-magick. For these tests, version 6.7.7.10 was used

[3] http://www.digitisation.eu/tools-resources/tools-for-text-digitisation/ncsr-border-detection-and-removal/

  - Abbyy FRE 11 Command Line Interface Version (CLI)
  - Abbyy FRE 12 Pro Version.

- OCR Evaluation tools:
  - ocrevalUAtion.[4]

These tools were executed in the form and with the parameters listed below:

- Image Magick deskewing: `convert input_file -deskew 40 output_file`
- App-deskew: `deskew64 -o output_file input_file`
- NCSR Border removal was executed through the web-service at Impact CoC Demonstrator Platform with option page split disabled.
- Image Magick border removal: `convert input_file -fuzz 59% -trim +repage output_file`[5]
- Image Magick binarisation: `convert input_file -colorspace gray +dither -colors 2 -normalize output_file`
- Abbyy FRE 10 binarisation: `AbbyyBinariser.exe input_file.tif output_file.tif params\params.ini` [params.ini is a text file containing license number and other data]
- Abbyy FRE 11 SDK for Dutch: `fr11-extended.jar -ocr -l Dutch -f TT_Typewriter -i input_file -e FEF_ALTO -o output_file`
- Abbyy FRE 11 SDK for Dutch and German: `fr11-extended.jar -ocr -l Dutch German -f TT_Typewriter -i input_file -e FEF_ALTO -o output_file`
- Abbyy FRE 11 SDK with user dictionaries: `fr11-extended.jar -ocr -l Dutch -f TT_Typewriter -w 50 -i input_file -e FEF_ALTO -o output_file -d userDictionary`
- Abbyy FRE 11 CLI for Dutch: `abbyyocr11 -rl Dutch -rtt Typewriter -f ALTO -if input_file -of output_file`
- Abbyy FRE 12 was run through its graphic interface with the following (alternative) options:
  - Document language: Dutch / Dutch & German

[4] http://www.digitisation.eu/training/succeed-training-materials/ocr-evaluation/ocrevaluation

[5] The optimal fuzz value when applied after App-dewkew tool is 59%. This parameter needs to be adjusted according to the set of images to which the tool is applied.

  - Document type: Typewriter
  - Color mode: Full color
  - Abbyy FRE 12 Image preprocessing: All options disabled / All options enabled but "split facing pages", "invert images" and "whiten background".

4 Methodology

The systematic comparison of the accuracy obtained under different settings requires the creation of ground truth transcriptions (transcriptions which are produced or revised manually to guarantee a 100% fidelity). To this purpose, the 89 documents of the collection were transcribed manually. The selection aimed to include documents with different characteristics, thus covering a wide variety of features appearing in the collection.

The transcription was made trying to follow the natural reading order, but since some documents present a complex-layout that allows for different reading orders, the results of the OCR were evaluated in word accuracy rate order-independent.

The average accuracy obtained with this test set can be only extrapolated with some care to estimate the accuracy on the full collection —even when identical processes and settings are applied—, since the global frequency of each feature may be slightly different. The variability of the results will be discussed in section 5.6.

The following operations were applied to every image in the test set:

- A deskewing tool was used to correct accidental rotations of the image.
- A border removal software allowed to remove empty areas surrounding the document.
- The binarisation of the images produced black and white versions with enhanced contrast (in particular, it whitens the background colour) for easier character recognition. Furthermore, binarisation often helps to remove artifacts produced by page transparency.[6]
- All images (the original and the enhanced ones) were processed with:
  - Abbyy FineReader Engine 11 SDK under four different settings: equipped only with its internal Dutch dictionary; equipped with its internal Dutch and German dictionaries; equipped with an external (user supplied) dictionary that contained a list of named

[6]The three preprocessing steps were applied in different orders to establish the optimal digitisation workflow.

entities (cities, streets and people); and equipped with the aforementioned external dictionary and applying the pattern training feature.
  - Abbyy FRE 11 CLI equipped with the internal Dutch dictionary.
  - Abbyy FRE 12 Pro equipped with the internal Dutch dictionary and several combinations of preprocessing options provided by this engine; with its internal Dutch and German dictionaries.

- The OCR output obtained was compared with the manual transcription in order to measure the accuracy and determine the optimal digitisation workflow for the collection using the ocrevalUAtion tool.

5 Results

The tools tested at every step, their specific settings and the results obtained are described below.

5.1 Deskewing

The performance on the test set of two deskewing tools has been evaluated:

- Image Magick deskewing tool.
- App-deskew from Galfar's Lair (http://galfar.vevb.net/wp/projects/deskew) available at Bitbucket (https://bitbucket.org/galfar/app-deskew).

Both tools provided good results when applied to the sample. However, App-deskew allows for a simpler automatic processing of the images. In contrast, Image Magick uses a threshold,[7] a parameter which must be adjusted for every particular image.

5.2 Border removal

Border removal tools crop the non-textual regions surrounding an image, in particular, background surfaces. Two border removal tools were applied to the images in the test set:

- NCSR Border removal
- Image Magick border removal

The best results were obtained when the images were surrounded by a black background and when Image Magick was applied. Some text was lost when the NCSR Border removal was used, while borders remained when Image Magick border removal was applied to images with white background.

[7] Image Magick deskewing can operate automatically, but the results were significantly worse.

5.3 Binarisation

Binarisation transforms a colour image into a black and white one and allows, for example, to reduce bleed-through. The tests showed that Abbyy FRE binarisation provides the best binarisation. The service used in the experiments was developed by the Impact project based on the Abbyy FRE 10 toolkit and updated by the Impact Centre of Competence to the Abbyy FRE 11 version. It does not require the adjustment of configurable parameters.

5.4 Determination of the order of the steps in the workflow

The order in the application of the tools which yielded the best results was the following:

1. Deskewing.
2. Border removal.
3. Binarisation.
4. OCR.

On the one hand, when border removal was applied before deskewing, some dark regions were not correctly identified and remained intact in the image. On the other hand, when the binarisation was applied before border removal, the binarisation became too aggressive and produced further loss of text. As it will be discussed in section 5.6, the experiments suggest that no binarisation at all produces the best results with this collection.

5.5 OCR engines and parameters

The automatic transcription of text was obtained with the Abbyy FineReader Engine 11 SDK, the latest version of the Abbyy Software Development Kit (SDK). The SDK is required to integrate user dictionaries (list of words which supplement the internal dictionary).

The experiments tested the performance of the FRE when the internal and the external dictionary are assigned equal relative weights, since the 50% combination provides optimal results according to the Impact Centre of Competence experience. The external dictionary contained 464,053 entries and it was built from two different sources:

- Gazetteers provided by the Instituut voor Nederlandse Lexicologie containing Dutch, Belgian, French and German locations, including river and sea names.
- German and Dutch names in the Geonames database (http://www.geonames.org).

The customizable parameters in Abbyy FRE 11 SDK were set as follows:

- Language: Dutch / Dutch and German
- Font: Typewriter
- Output format: ALTO

Abbyy FRE 11 SDK allows for pattern training[8] although, according to the manufacturer, this is only recommended for text sets in decorative fonts, texts containing unusual characters or long documents (over one hundred pages) of low print quality. The feasibility of such fine-tuning with a limited number of pages was explored but, as shown in Table 3, a higher error rate was found after training.

In order to compare the accuracy of the character recognition with different versions of the Abbyy FRE, tests were also performed with the Abbyy FRE 11 CLI and with the Abbyy FRE 12 Pro. In contrast to the SDK version, the CLI and Pro versions do not allow for the integration of external dictionaries. The customizable parameters in Abbyy FRE 11 CLI were set as follows:

- Language: Dutch
- Font: Typewriter
- Output format: ALTO

The customizable parameters used in Abbyy 12 Pro were:

- Language: Dutch / Dutch and German
- Font: Typewriter
- Output format: raw text (TXT)
- Colour mode: Full colour
- Abbyy FRE 12 Image preprocessing: All options disabled / All options enabled excluding "split facing pages", "invert images" and "whiten background".

5.6 Analysis of the results

The results were evaluated with the ocrevalUAtion tool by measuring the order-independent word accuracy with respect to the ground truth transcriptions. Strict ordering of the words was not enforced because the complex layout of the documents often allows for more than one correct reading order of the text regions in the document. Punctuation errors were also ignored. Table 1 summarises the accuracy under different versions of Abbyy FRE, settings and combinations of preprocessing steps.

[8] See https://abbyy.technology/en:features:ocr:pattern_training.

	No preprocessing	Deskewing	Deskewing & Border removal	Deskewing & Border removal & Binarisation
FRE 11 SDK	73.1	79.9	81.1	79.5
FRE 11 SDK + User dictionaries	64.1	71.9	72.3	70.7
FRE 11 SDK + User dictionaries + Pattern training	63.7	71.9	71.7	70.0
FR12 Pro - No Abbyy Preprocessing - Dutch & German	76.3	78.4	79.2	77.8
F12 Pro - Abbyy Preprocessing - Dutch & German	70.7	77.5	77.4	77.3

Table 1: Word accuracy when no preprocessing takes place (1st column), when deskewing is applied (2nd column), when border removal and deskewing are applied (3rd column), and when all preprocessing steps (deskewing, border removal and binarisation) are applied (4th column).

The effect of preprocessing. Results show that deskewing and border removal tools improve the accuracy of the OCR. However, binarisation leads to higher error rates because it produces loss of text when applied to this kind of material: while ink density is usually homogeneous in printed documents, this collection contains mainly typewritten content where the characters show variable levels of darkness. Since lighter parts tend to disappear in the binarisation process, this often leads to a lower accuracy.

The influence of dictionaries. The accuracy obtained when Abbyy FRE 11 SDK is equipped with both Dutch and German dictionaries turned out to be analogous to that obtained when only the Dutch dictionary is selected. In contrast to past experiences, the use of external dictionaries did not improve in this case the accuracy of the results, as revealed by the comparison of rows 1 and 2 in table 1.[9] As it was expected, pattern training did not lead to higher recognition rates (as seen in table 1, row 3).

The effect of a manual pre-classification of the documents according to their main language and the selection the corresponding internal dictionary) has not been evaluated.

Dependency on the Abbyy FR version. Experiments with Abbyy FRE 11 CLI were performed in order to check that the results were analo-

[9] Unfortunately, the closed nature of FR does not allow to explain the exact reasons for the observed behaviour.

	No preprocessing	Deskewing	Deskewing & Border removal	Deskewing & Border removal & Binarisation
Black background	73.1	79.9	81.1	79.5
Dark Grey background	72.5	80.4	80.4	80.0
Light Grey background	73.1	78.8	78.7	78.5
White background	69.5	77.5	77.2	76.5
Physical folder	57.6	64.5	66.0	64.9

Table 2: Word accuracy when no preprocessing takes place (1st column), when deskewing is applied (2nd column), when border removal and deskewing are applied (3rd column), and when all preprocessing steps (deskewing, border removal and binarisation) are applied (4th column).

gous to those obtained with the SDK version. It was found that the error rate slightly raised with respect to the SDK[10]. For example, the accuracy decreased from 81.1% to 80.8% when the images were processed with the FRE 11 CLI after border removal and deskewing.

Tests were also performed with the latest version of Abbyy FRE Pro (currently, number 12). The FRE 12 Pro worked best when no Abbyy preprocessing step is applied. No significant difference was observed (compare the fourth and fifth rows in table 1) between the Dutch+German language option and the Dutch only are not significant. In all cases, the accuracy was lower than that obtained with the SDK version.

Variations in the background. The results were evaluated with images placed on a white background, on a black background and with images surrounded by the context where they were found in the containing physical folder (for example, other documents) and they are summarized in table 2. While there are only minor differences between using a black or a white background, the images in context lead to a much lower accuracy.

Variability of the results. Figure 3 presents a histogram showing the frequency of the accuracies obtained after the documents in the sample are processed with the proposed workflow. The transcription of most documents reaches an accuracy rate over 80%, but there is a small fraction of documents where the accuracy falls below 10%. Although documents with blurred letters were initially considered to be more difficult for OCR, the inspection of the results proved that documents with lowest accuracy are mainly cards or documents with very complex layout. This fact may be taken into account, for example, to perform a manual pre-classification of the documents or to

[10]Private communication with Abbyy did not revealed the reason for this difference

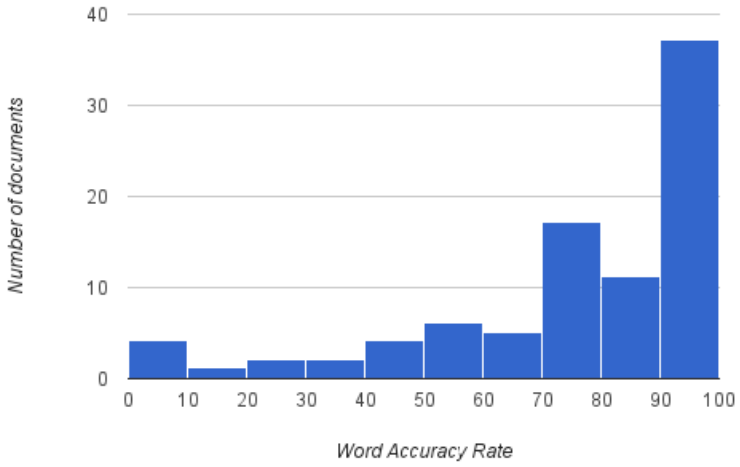

Figure 2: Distribution of word accuracies (X = Word Accuracy Rate Y = number of documents.

define a specific workflow for this type of images.

5.7 Summary of results

- Image preprocessing tools work best when applied on images placed on black background.
- The automatic transcription works best when deskewing and border removal are applied but no binarisation process takes place.
- The use of external dictionaries does not lead to a higher accuracy in the OCR process.
- The Abbyy FR version which provides the best results when applied to the current collection is the Abbyy FR11 SDK.

6 Conclusions

The comparative analysis of different settings and workflows for the automatic transcription of the collection described in this report leads to the conclusion that the best results are obtained when the following steps (recommended workflow) are applied:

1. Deskew images using the Deskew Tool by Galfar's Lair. This is a command line tool which can be simply invoked as follows:

```
deskew64 -o output_file input_file
```

2. Crop empty regions around text with the Image Magick Border Removal:

   ```
   convert input_file -fuzz 59% -trim +repage output_file
   ```

3. Perform OCR with Abbyy FRE 11 SDK with the following options:

   ```
   fr11-extended.jar -ocr -l Dutch -f TT_Typewriter -i input_file
   -e FEF_ALTO -o output_file
   ```

A Selection and specification of the test-set (CABR pilot)

The selection is such that it has a large variety of different documents. The documents in this test are selected from 2 different inventory numbers.

The documents in the selection contain:

- Coloured paper
- Both typed and hand written text on a page
- Faded text
- Text in columns
- Different colours carbon paper
- Different paper sizes
- Typed names
- Typed names with extra spacing
- Typed names with underlining
- Broken names
- Mirrored text at the back of the paper visible
- Stencils
- Blue text
- Purple text
- White text with a black background (photographs)
- Blurry text
- Clear text
- Stencils with columns and quotation marks
- Text with printed line dots (.....)
- Text in italics
- Ornamental letters
- Photocopies
- Small letters

- Printed, typed and handwritten text
- Columns with dates
- Messy text
- Transparent paper
- Copying paper
- Partially deleted text
- Typed, printed and colour stamps
- The upper side of the typed letters fall away
- Strikethrough text
- Typed list and dots
- Forms
- Coloured cards
- Text cut off

The scans are made in TIFF 6.0. There is no post processing used. The converted JPEG files have a baseline compression 1:10.

For the requirements and image quality see the attached documents:

- 2._Offerteaanvraag Digitalisering_EOI.CABR.pdf
- AANB DIGI_BESCHRIJVEND DOCUMENT_DEF.pdf
- BIJLAGE J EISEN_DEF.xls (zie de eisen voor het kavel: archieven, standaard)

The digital images are made with different backgrounds to determine which background gives the best OCR results. The used backgrounds for the documents are:

- White
- Black
- Light grey
- Dark grey
- In context (stack of papers)