Lexicon-supported OCR of eighteenth century Dutch books: a case study

Jesse de Does[a] and Katrien Depuydt[a]

[a] INL, Postbus 9515, 2300 RA, Leiden, Netherlands

ABSTRACT

We report on a case study on OCR of eighteenth century books conducted in the IMPACT project. After introducing the IMPACT project and its approach to lexicon building and deployment, we zoom in to the application of IMPACT tools and data to the Dutch EDBO collection. The results are exemplified by detailed discussion of various practical options to improve text recognition beyond a baseline of running an uncustomized Finereader 10. In particular, we discuss improved recognition of long *s*.

Keywords: OCR, historical language, historical lexicon, document recognition, evaluation, text recognition.

1. INTRODUCTION

1.1 The IMPACT project

IMPACT (2008-2012) is a project funded by the European Commission. Its aim was to significantly improve access to historical text and to take away the barriers that stand in the way of the mass digitization of the European cultural heritage. For that purpose IMPACT aimed to improve the quality of OCR (Optical Character Recognition) for historical documents and to enhance their accessibility. The project consortium consisted of 26 partners (eleven libraries, thirteen research institutes or universities, and two private sector companies, ABBYY and IBM). The project is followed up by a Centre of Competence (www.digitisation.eu) aimed not only to make available the results from IMPACT but also to build a sustainable environment that will allow research institutes, private sector partners and cultural heritage organisations to work together to continue to improve access to historical texts[1].

There are many aspects involved in dealing with the problems addressed by IMPACT. Image processing, which tries to remedy typical problems like skewed, warped or otherwise noisy data; better segmentation procedures and adaptive OCR aim to overcome the irregularities of historical typography. The present contribution focuses on enhancement of OCR results by using the appropriate historical lexica.

1.3 Lexicon building and deployment in IMPACT

Full-text accessibility for historical text documents is hindered by the “historical language barrier”, caused by historical spelling and language variation. The language is obviously an issue in full-text retrieval. In IMPACT lexica[2] have been built enabling users to access digitised texts without having to take into account all possible spellings and inflections of words.

Historical language and spelling variation is also an issue for OCR. For good quality OCR linguistic resources such as lexica and language models are important. Language models are not in the scope of this paper. Modern OCR engines use background lexica for distinct languages to weigh, verify and correct preliminary recognition results of the symbol

[1] http://www.digitisation.eu/about/.

[2] In IMPACT improving access to text means not only improving OCR, but also improving retrieval. For the latter, improvement of OCR is a first step. In addition to that, IR lexica were built for each language, enabling users to find words in historical spelling by using the modern spelling.

classifiers. For optimal effect, it is important that the lexicon used covers the vocabulary of the input text and that the spelling in the lexicon and in the input text coincide. This implies that special lexica are needed for historical lexica. One achievement of the IMPACT project is the development of OCR lexica for historical variants of nine European languages[3].

The development of lexica (whether for IR or for OCR) cannot be undertaken successfully without considering the options for deployment of the data. The historical OCR lexica developed in the project have been deployed with both the FineReader engine and with the IMPACT adaptive OCR engine[4]. However, the contribution to OCR quality obtained by use of special lexica has only been measured with FineReader. For a full report, cf. the "Cross Language Perspective"[5] on OCR and IR results using historical lexica.

1.3 IMPACT lexicon deployment in OCR

One option for applying the OCR lexica produced in IMPACT would have been to extend FineReader (or any other OCR engine) by building in these lexica in the same way lexica for modern languages are implemented by the engine. However, all such dictionaries are built and maintained by ABBYY, and presently there are no externally available tools for building FineReader dictionaries. Hence, we had to resort to a different approach. The FineReader Engine SDK has an interface for binding so-called "external dictionaries". This interface has been improved by ABBYY in the course of the project. We implemented this interface in order to conduct the evaluation experiments in which OCR lexica with historical vocabulary for all IMPACT languages were used to improve FineReader.

Use of the FineReader external dictionary interface in IMPACT

A usable implementation of the FineReader external dictionary interface requires:

1. Implementation of a C++/COM class interface. Briefly, this consists of implementing two methods:
   a. A method which prunes a "fuzzy set" of word recognition candidates to the subset of linguistically valid ones, providing each valid recognition candidate with a confidence score between 1 and 100.
   b. A method which decides whether a set of recognition candidates contains a prefix which can be extended to a valid word.
2. Development of a simple FineReader SDK-based OCR-executable application which actually uses this implementation during recognition.

The IMPACT implementation, which will be made available in the Center of Competence[6], consists of:

1. The definition of a "plain C" version of the external dictionary interface, and the development of a Windows DLL implementing this plain C interface, using (a binary compilation of) a static word list with "confidence" information to prune and weigh recognition candidates.
2. An executable which is an adapted version of the CommandLineInterface SDK demonstration program which is part of the FineReader engine distribution. The executable implements the External Dictionary Interface by proxy: the actual work is done in the dynamically loaded DLL module, which is specified on the command line.
3. A small utility program to compile a word list with scores to the required binary format required by 1).

[3] Dutch, German, English, French, Spanish, Polish, Czech, Bulgarian and Slovene. For information on the toolbox for lexicon building and deployment, http://www.digitisation.eu/tools/toolbox-for-lexicon-building/. For information on the language resources, http://www.digitisation.eu/tools/language-resources/.

[4] Advancement of OCR has been pursued in two distinct ways: by enhancing the (leading) FineReader OCR engine in various ways on the one hand, and by the development of the IBM adaptive OCR engine on the other.

In the project, FineReader has been enhanced, for instance, by improving support for Gothic fonts, improved binarization and improved support for "external" lexical data. The IBM engine focuses on adaptivity. It is tightly coupled to the innovative CONCERT tool for interactive post-correction. For further information, see www.digitisation.eu.

[5] D-EE2.8: *Use of Computational Lexica for OCR and IR on historical documents – a cross-language perspective.* To be made available at www.digitisation.eu.

[6] http://www.digitisation.eu.

The DLL has been used by Content Conversion Specialists GmbH to test the effect of the Dutch historical IMPACT lexica in an actual OCR workflow by means of integration in the docWorks Large Scale Digitisation Workflow system[7]. As a result of this test, the lexicon (including the software) has been purchased and used in the digitisation of Dutch historical newspapers.

The results in this paper have been obtained using Finereader engine version (version 10, build 10.0.3.494).

1.4 Cross-language evaluation of lexicon-supported OCR results in IMPACT

An extensive evaluation of the contribution of the IMPACT lexica to text recognition has been conducted. The evaluation was carried out by comparing FineReader Engine version 10 in its optimal internal dictionary and language setting (for English, French, German and Spanish, already available historical dictionaries in the FineReader SDK distribution have been used) with FineReader using the same internal dictionary combined with an external historical dictionary that was run through the FineReader external dictionary. A purely scientific comparison between the current internal lexica and the external IMPACT lexica was not feasible, because we lack information on how the dictionary is used internally, and are not able to convert the IMPACT lexica into the internal format. Moreover, our main concern was and is to enhance existing functionality rather than replacing it.

The results are based on a word-based, case- and punctuation-insensitive alignment of ground truth and OCR. Since the alignment is done region-by-region, complex layouts can still be more or less evaluated. We developed a custom evaluation tool that enables us to use the layout and coordinate information in the IMPACT ground truth XML files, and to obtain more detailed statistics on for instance frequent word errors, dictionary word hallucinations, dictionary coverage, …..

The evaluation data for each IMPACT language (Bulgarian, Czech, Dutch, English, French, German, Polish, Slovene, Spanish) consist of a random selection of about 200 pages from the “Evaluation” subset of the ground truth transcriptions[8]. The following chart summarizes results on the evaluation sets.

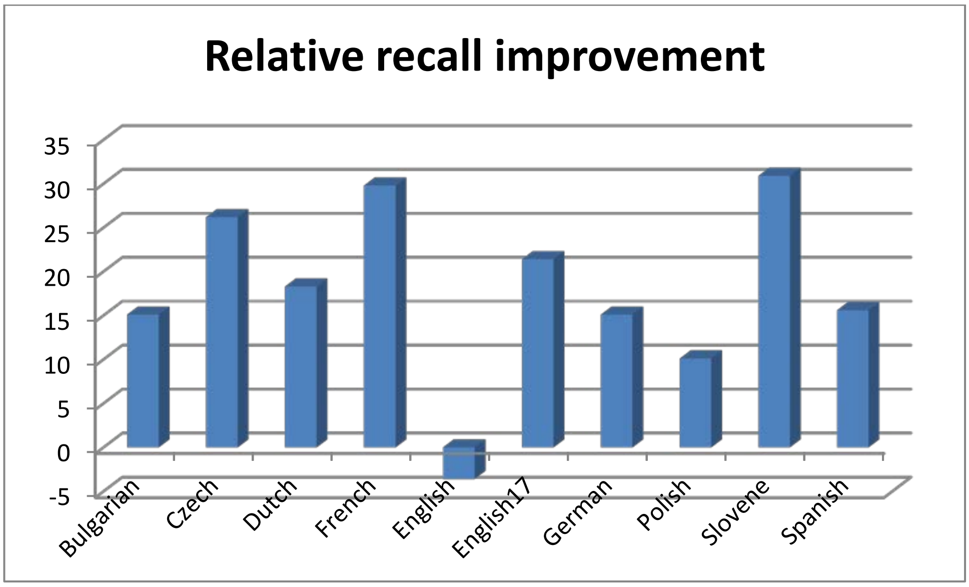


Figure 1: relative word recall improvement of OCR using historical lexica

[7] Presentation at 2011 IMPACT conference, (Gravenhorst, 2011, http://vimeo.com/31999737).

[8] http://www.digitisation.eu/tools/research-dataset/ for more information on the dataset from which the evaluation pages were selected.

We cannot go into full detail about the lessons learned from these experiments. We summarize some observations below.

- The assumption that adding a type frequency list without further cleaning would suffice as a good OCR lexicon proves to be wrong. Even if good lexica and ground-truth quality corpora are available, the development of an OCR lexicon is not a completely obvious task.
- *Relevance* of lexical data is obviously important, but the degree to which this is borne out by the experiments varies. On the one hand, we find that using a general-purpose historical lexicon for English only deteriorates results, whereas a focused 17th century lexicon does contribute to OCR quality. On the other hand, a general-purpose Dutch historical lexicon achieves good results.
- Clearly, a significant degree of coverage on the target material is necessary. On the other hand, including rare and unusual words in the lexicon of accepted words may lead to so-called "dictionary hallucinations", where the OCR engine "corrects" a perfectly valid word to a word form from the lexicon (a "false friend").When extracting an OCR lexicon from a large corpus, it does not make sense to include low-frequency words (below frequency 5 or 3).
- Similarly, short words (length 5 and lower) should only be included to the point that a certain degree of corpus coverage, as measured within the set of words of a certain fixed length, has been reached[9]. The reason is that unfrequent short words appear to often contribute more to dictionary hallucinations than to correctly recognized words.
- When a lower frequency word is related to a high frequency word by a frequent OCR confusion, it often improves performance to omit the lower frequency word (this may for instance lead to exclusion of words like *fecond* and *fur* from a French OCR lexicon for documents using long s).

1.5 The Dutch IMPACT lexicon for OCR and Retrieval

The Dutch IMPACT lexicon[10] is intended to improve both OCR and retrieval for historical Dutch Documents. It implements the main characteristics of IMPACT lexica. All word forms in the retrieval lexicon are provided with modern lemma and main part of speech. An important feature is the inclusion, for all words forms described in the lexicon, of dated attestations, permitting the extraction of period-specific sublexica.

The Core General lexicon lexicon for Dutch relies on the following data:

1. The result of dictionary-quotation-based attestation from the Woordenboek der Nederlandsche Taal (WNT[11])

2. The result of corpus-based lexicon building from a selection of the Dutch "DBNL"[12] historical corpus material

3. The result of corpus-based lexicon building from selections of KB Material (morphological module)

The material ranges from 1550 to 1970, thus providing a core around which more specific lexicon data based on selected corpora can be developed The IR lexicon currently contains 475498 distinct word forms, 215180 lemmata, 558438 distinct lemma/wordform combinations.

2. IMPROVING OCR RESULTS FOR THE "EDBO" COLLECTION

After the termination of the IMPACT project proper (end of 2011), the project was extended for six months to enable a few pilot studies to test IMPACT tools.

The aim of these "second extension pilots" was to investigate real-life situations and put IMPACT tools to the test. In this context, the National Library of the Netherlands (KB) has tested options for improving the OCR results of the

[9] We used 99% with good results for Dutch and seventeenth-century English and Spanish

[10] Available from INL and center of competence, cf http://www.digitisation.eu/tools/language-resources/historical-lexicon-dutch/

[11] Cf. for instance Mooijaart 2010[23]. The online version of the dictionary in combination with three other major historical dictionaries is at http://gtb.inl.nl

[12] Digitale Bibliotheek voor de Nederlandse Letteren, Digital Library for Dutch Literature, www.dbnl.org

existing "Early Dutch Books Online" collection[13]. The main part of their pilot study is devoted to testing the post-correction tools developed in IMPACT, though re-OCR'ing with the support of the Dutch historical dictionary is also discussed. As a companion to the KB study, we have investigated the OCR for these materials in more detail. Below we will report on our findings and discuss practical options for improving results without manual intervention.

2.1 Options for improving text quality

The main objective of this study is to assess practical options for the deployment of historical lexical data to improve text recognition. Apart from approaches involving manual correction of the collection, the main options to improve the text quality are re-OCRing with an improved OCR setup and automatic post-correction.

Improving OCR

The main obstacles for good OCR of historical documents - apart from image quality, which is not the most pressing problem in the EDBO collection - are historical typography and historical language. Accordingly, we have considered two options to improve the OCR process itself:

1. OCR with improved settings for recognition of the historical Dutch character set
2. Lexicon-supported OCR with IMPACT historical lexicon for Dutch

Post-correction

Since re-OCR-ing is not always practically (or financially) achievable, the question naturally comes up how much can be achieved by fully automatic post-correction. Both approaches to post-correction developed in IMPACT (the CONCERT Tool[14] developed by IBM and the post-correction tool[15] developed by LMU) are delivered as interactive tools[16]. Hence the last option considered:

3. Naïve automatic post-correction using the with IMPACT historical lexicon (more detail in section 2.1.3)

As we shall see, a huge proportion of the errors in the baseline OCR results from the confusion of long s with f. Accordingly, our implementation of options 1-3 for this paper is targeted to the solution of this particular problem. More detail is given below.

2.1.1 Baseline

As a baseline, we have run Finereader 10 with default settings for the Dutch language. The results in this paper have been obtained using Finereader engine version (version 10, build 10.0.3.494).

2.1.2 Improving settings for historical Dutch: customization of character set

The Finereader engine has the option of adding long s (or any other glyph supported by the engine) to the recognition character set by means of an API call. We re-ocred the two books with this option, without other changes to the baseline setup.

[13] http://www.earlydutchbooksonline.nl/nl/edbo. The OCR used for the EDBO website has been produced by means of Finereader 9.0. The OCR quality has been evaluated in the KB pilot for the two books under consideration in this study. Accuracy is reportedly 84%.

[14] http://www.digitisation.eu/tools/ocr-post-correction-and-enrichment/collaborative-correction-platform/

[15] http://www.digitisation.eu/tools/ocr-post-correction-and-enrichment/post-correction-tool/

[16] The IBM approach to post-correction is inherently interactive. An advanced approach to automatic error detection and suggestion of correction candidates has been developed and implemented by LMU. However, partly as a response to the apprehensiveness of libraries to implement automatic post-correction which might also damage correctly recognized words (as is inevitable for any fully automatic approach to post-correction), it was preferred to deliver this post-correction system as an interactive tool.

2.1.2 Lexicon-supported OCR with “Long s fix”

The external dictionary implementation (cf. 1.3) contains a workaround for a frequent problem in OCR of historical documents: the recognition of long s vs. f. Even when long s is added to the FineReader character set[17], differentiating the two remains problematic. One of our findings is that it is not always an option to relegate this to the post-correction stage, as the s/f problem may cause the engine to output a completely different recognition candidate, which may be beyond repair by post-correction. For instance Dutch *eerste* (first) is turned into/misrepresented as the dictionary word *eerde (*“honoured”). By having the external dictionary basically “accept” the alternative recognition candidate *eerfte* and correcting it to *eerste* before output, it turns out we can improve recognition.

2.1.3 Naïve post-correction of the baseline OCR

For this case study, we have merely implemented the most straightforward form of post-correction for the f/s problem. Any word *w* not in the historical lexicon, which can be transformed into a lexicon word w_1 by substituting one or more instances of *f* by *s*, is replaced by w_1.

2.2 Experiment setup

2.2.1 Data

We have used the two eighteenth century books from the Dutch *early dutch books online* collection selected by the KB for their pilot.

1. Verzameling van placaaten, resolutien en andere authentyke stukken enz. betrekking hebbende tot de gewigtige gebeurtenissen, in de maand september MDCCLXXXVII, bevooren en vervolgens, in het gemeenebest der Vereenigde Nederlanden voorgevallen. : Part 29.

   Year: 1791

   Printer/publisher: Chalmot, Jacques Alexandre de Kampen, 1778-1797

   Copy: Leiden, Universiteitsbibliotheek: 1006 A 29

   No. of pages: 338

   EDBO URL: http://www.earlydutchbooksonline.nl/nl/view/image/id/dpo:3077:mpeg21

2. Verhandelingen van het Genootschap ter bevordering der heelkunde, te Amsterdam. : Part 1

   Year: 1791

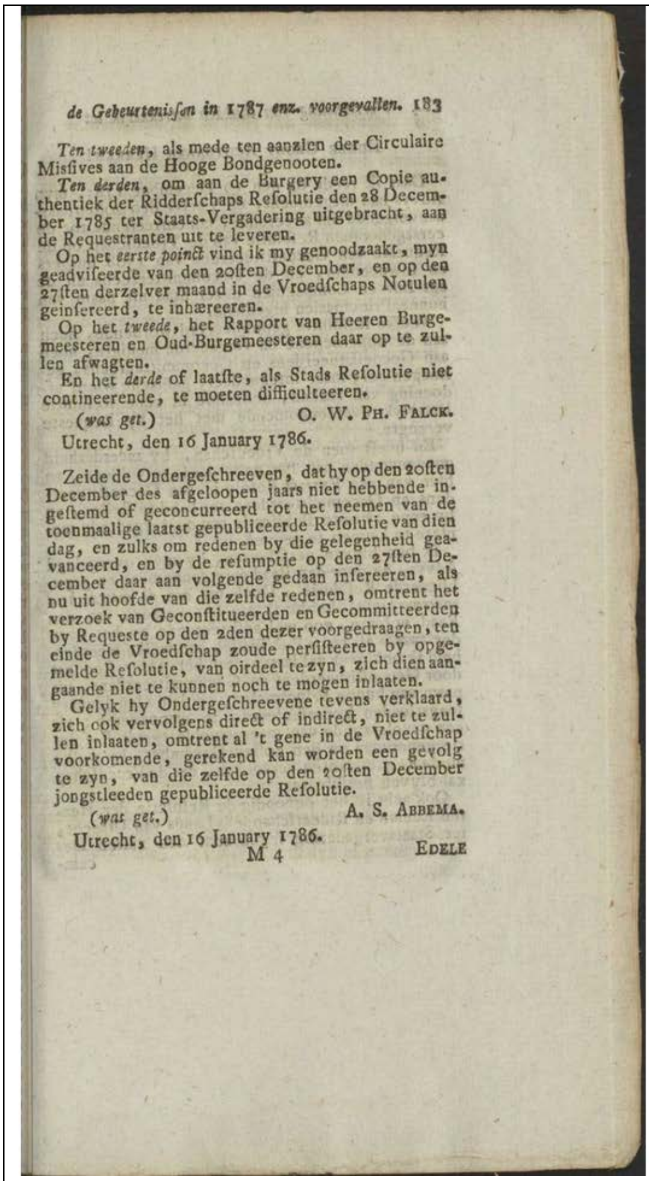
   Printer/publisher: Elwe, Jan Barend Amsterdam, 1778-1800

   Copy: Leiden, Universiteitsbibliotheek: 1448 G 2

   No. of pages: 331

   EDBO URL: http://www.earlydutchbooksonline.nl/nl/view/image/id/dpo:3423:mpeg21

[17] It is already included in the default character set for languages with an “Old” dictionary (English, French, German, Spanish). There is an option in the SDK to alter the set of accepted characters.



de Gebeurteniſſen in 1787 *enz. voorgevallen.* 183

Ten tweeden, als mede ten aanzien der Circulaire Misſives aan de Hooge Bondgenooten.

Ten derden, om aan de Burgery een Copie authentiek der Ridderſchaps Reſolutie den 28 December 1785 ter Staats-Vergadering uitgebracht, aan de Requestranten uit te leveren.

Op het *eerste poinct* vind ik my genoodzaakt, myn geadviſeerde van den 20ſten December, en op den 27ſten derzelver maand in de Vroedſchaps Notulen geinſereerd, te inhæreeren.

Op het *tweede*, het Rapport van Heeren Burgemeesteren en Oud-Burgemeesteren daar op te zullen afwagten.

En het *derde* of laatſte, als Stads Reſolutie niet contineerende, te moeten difficulteeren.

(*was get.*) O. W. Ph. Falck.

Utrecht, den 16 January 1786.

Zeide de Ondergeſchreeven, dat hy op den 20ſten December des afgeloopen jaars niet hebbende ingeſtemd of geconcurreerd tot het neemen van de toenmaalige laatst gepubliceerde Reſolutie van dien dag, en zulks om redenen by die gelegenheid geavanceerd, en by de reſumptie op den 27ſten December daar aan volgende gedaan inſereeren, als nu uit hoofde van die zelfde redenen, omtrent het verzoek van Geconſtitueerden en Gecommitteerden by Requeste op den 2den dezer voorgedraagen, ten einde de Vroedſchap zoude perſiſteeren by opgemelde Reſolutie, van oirdeel te zyn, zich dien aangaande niet te kunnen noch te mogen inlaaten.

Gelyk hy Ondergeſchreevene tevens verklaard, zich ook vervolgens direct of indirect, niet te zullen inlaaten, omtrent al 't gene in de Vroedſchap voorkomende, gerekend kan worden een gevolg te zyn, van die zelfde op den 20ſten December jongstleeden gepubliceerde Reſolutie.

(*was get.*) A. S. Abbema.

Utrecht, den 16 January 1786.

M 4 Edele

OCR:

de Gebeurtenisje in 1787 enz. voorgevallen. ï.83

Ten tweeden, als mede ten aanzien der Circulaire Misfives aan de Hooge Bondgenooten. .

Ten derden, om aan de burgery een Copie au. thentiek der Ridderfchaps Refolutie den 28 December 1785 ter Staats-Vergadering uitgebracht, aan de Requestranten uit te leveren.

Oo het eerste poinct vind ik my genoodzaakt, myn aeadvifeerde van den 2often December, en op den i7ften derzelver maand in de Vroedfchaps Notulen geinfereerd, te inhajreeren.

Op het tweede, het Rapport van Heeren Burgemeesteren en Oud-Burgemeesteren daar op te zul-lSV

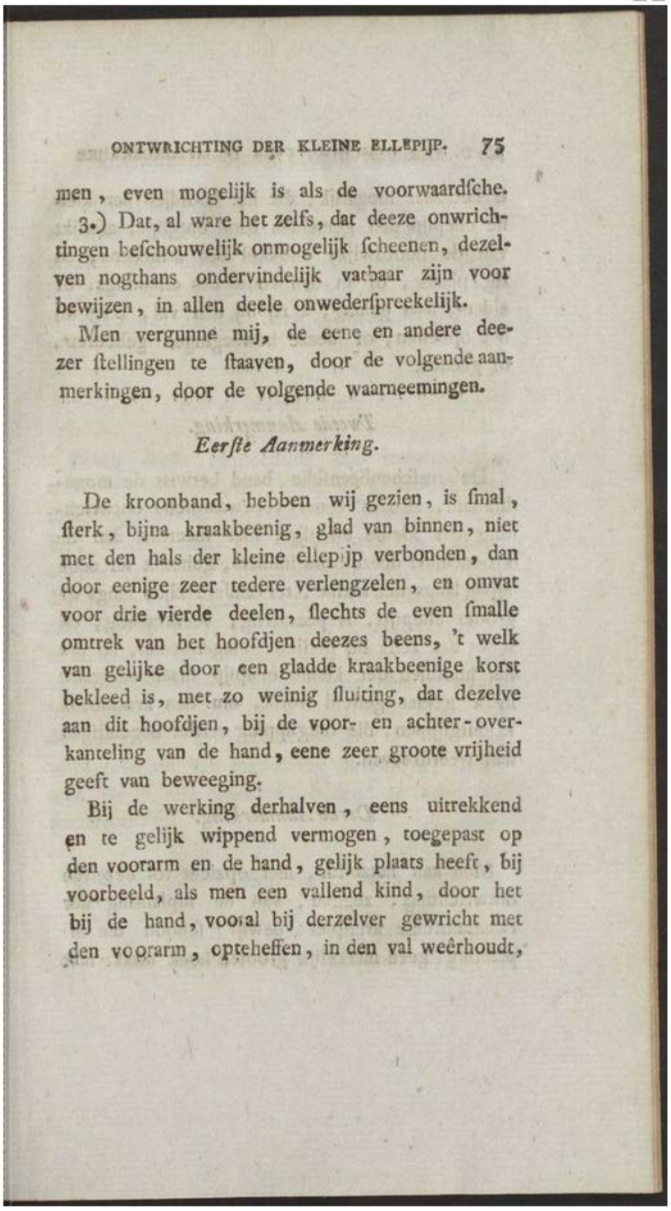
lhTderde of laatfte, als Stads Refolutie niet contineerende, te moeten difficulteeren.

(was g«.) O. W. Ph. Falck.

Utrecht, den 16 January 1786.

......

Figure 2: Book 1: Verzameling van placaaten, resolutien en andere authentyke stukken enz …



ONTWRICHTING DER KLEINE ELLEPIJP. 75

men, even mogelijk is als de voorwaardſche.

3.) Dat, al ware het zelfs, dat deeze onwrichtingen beſchouwelijk onmogelijk ſcheenen, dezelven nogthans ondervindelijk vatbaar zijn voor bewijzen, in allen deele onwederſpreekelijk.

Men vergunne mij, de eene en andere deezer ſtellingen te ſtaaven, door de volgende aanmerkingen, door de volgende waarneemingen.

Eerſte Aanmerking.

De kroonband, hebben wij gezien, is ſmal, ſterk, bijna kraakbeenig, glad van binnen, niet met den hals der kleine ellepijp verbonden, dan door eenige zeer tedere verlengzelen, en omvat voor drie vierde deelen, ſlechts de even ſmalle omtrek van het hoofdjen deezes beens, 't welk van gelijke door een gladde kraakbeenige korst bekleed is, met zo weinig ſluiting, dat dezelve aan dit hoofdjen, bij de voor- en achter-overkanteling van de hand, eene zeer groote vrijheid geeft van beweeging.

Bij de werking derhalven, eens uitrekkend en te gelijk wippend vermogen, toegepast op den voorarm en de hand, gelijk plaats heeft, bij voorbeeld, als men een vallend kind, door het bij de hand, vooral bij derzelver gewricht met den voorarm, opteheffen, in den val weêrhoudt,

OCR:

ONTWRICHTING DER KLEINE ELLEPIJP. f5

men , even mogelijk is als de voorwaardfche.

3.) Dat, al ware het zelfs, dat deeze onwrich tingen befchouwelijk onmogelijk fcheentn, dezel ven nogthans ondervjndelijk vatbaar zijn voor bewijzen, in allen deele onwederfpreekelijk.

Men vergunne mij, de eene en andere dee zer Hellingen te ftaaven, door de volgende aan merkingen, door de volgende waarneemingen.

Eer/te Aanmerking.

De kroonband, hebben wij gezien, is fmal, fterk, bijna kraakbeenig, glad van binnen, niet …

Figure 3: Book 2: Verhandelingen van het Genootschap ter bevordering der heelkunde, te Amsterdam

2.2.2 Metrics and Evaluation method

We have used the OCR evaluation tool as described in section 1.4. The version of the tool that was used in this report gives precision and recall for case-insensitive word accuracy, not counting punctuation, as the main evaluation metrics. Precision and recall, computed after region-by-region word-level alignment of OCR and ground truth, are defined as:

$$precision = \frac{\text{number of correctly recognized words}}{\text{number of recognized words (= number of words in OCR output file)}}$$

$$recall = \frac{\text{number of correctly recognized words}}{\text{number of ground truth words}}$$

For most normal pages, precision and recall are close to each other. Discrepancies arise when complete regions are erroneously detected as text (low precision) or image region (low recall).

2.3 Results

2.3.1 Baseline results

The baseline OCR has been produced by Finereader version 10, build 10.0.3.494, with default settings for Dutch.

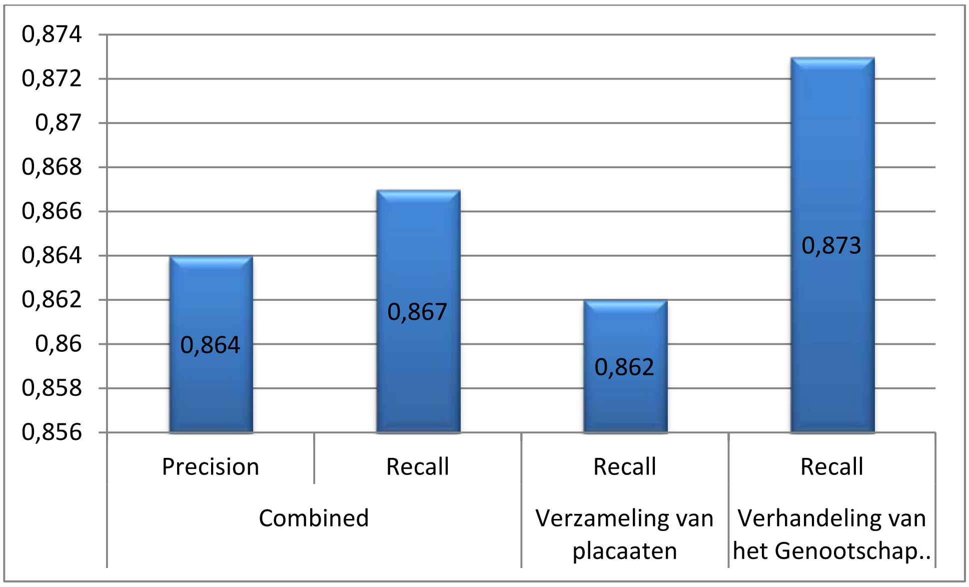


Figure 4: baseline OCR results

Typical errors in the baseline OCR

The following charts list frequent error frequencies in single word errors from the baseline OCR of the two books.

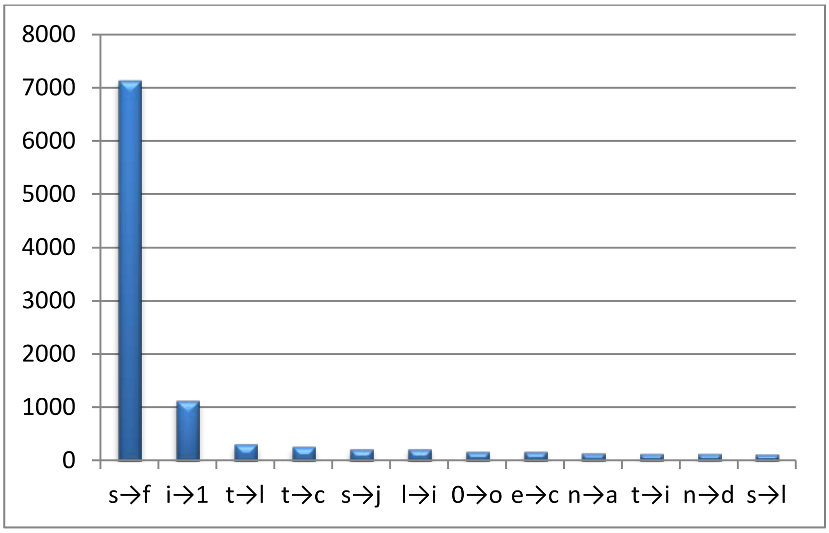


Figure 5: Frequent error types without extra lexicon: (in total about 16000 confusions)

As the table shows, a large part of the errors is due to a few frequent confusions. The most prominent error (f/s confusion) is especially obnoxious, as it causes misrecognitions of words to be more frequent than correct transcriptions in many cases. Cf. the following result for a lucene fuzzy query run in the BlackLab retrieval demonstrator[18] on the complete EDBO set for the keyword "*schoonheid*" (beauty):

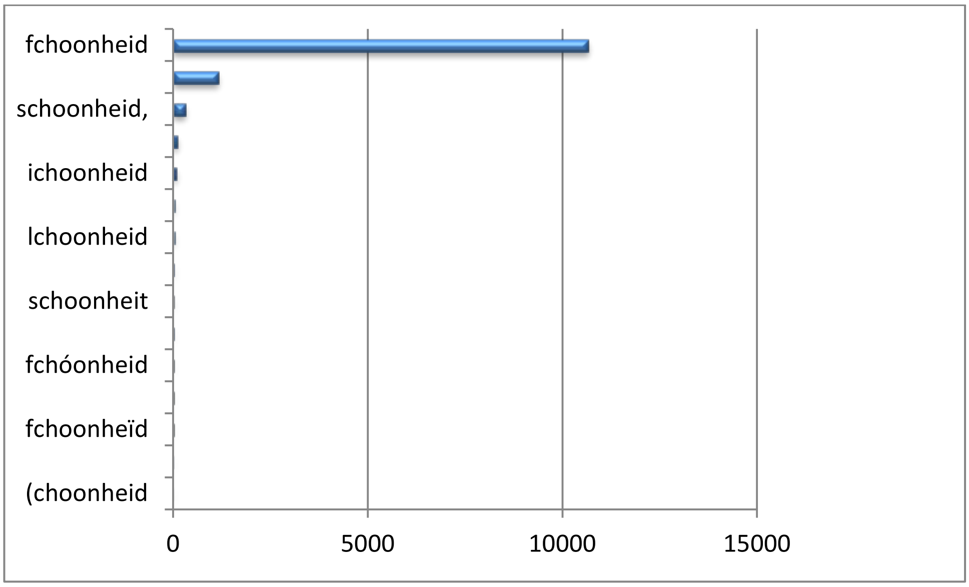


Figure 6: distribution of variant recognitions of "schoonheid" (beauty). "schoonheid" and "schoonheit" are correct, the other forms are OCR errors

2.3.2 Results with improved settings for historical Dutch

As it turns out, this simple option already has a significant effect on the quality of the text recognition. Precision increases to 89%, recall to 89.3%. Much to our surprise, service providers digitizing large collections for libraries tend not to apply this simple option.

2.3.3 Using the Dutch historical lexicon

With added historical lexicon (cf. section 1.5) we obtain a precision of 91.9% and a recall of 92.4%.

[18] To demonstrate the use of IMPACT lexica in retrieval, a lucene-based search engine (BlackLab), enabling the use of lexica and linguistic annotation, has been developed in the project. Available as open source, https://github.com/INL/BlackLab

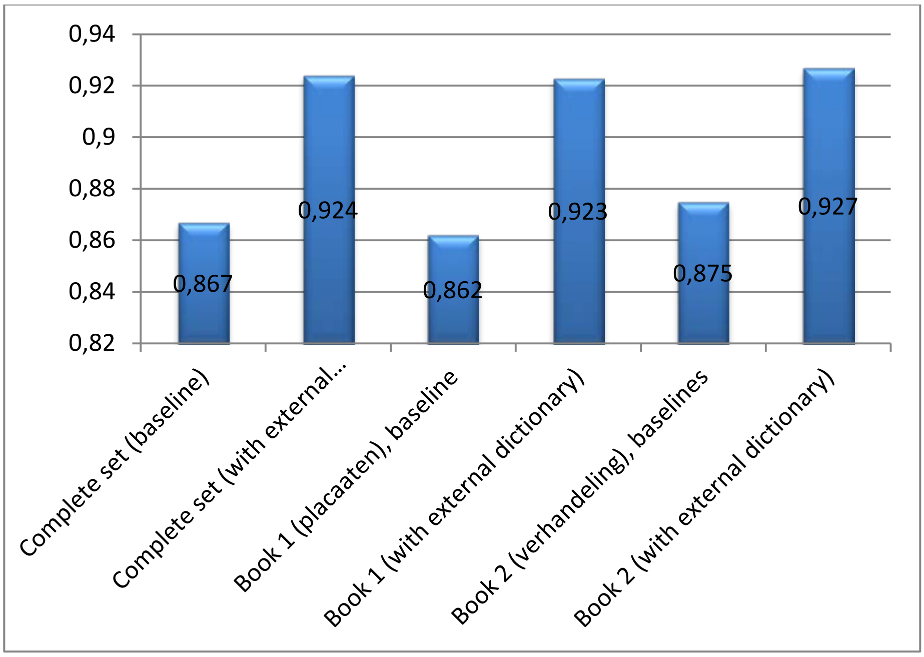


Figure 7: benefit of historical lexicon

It should be noted that combining the previous options (i.e using both the external dictionary and the customized character set) does not lead to significant improvement: precision: 0.921, recall: 0.926. The reason for this is that by adding the historical dictionary using the external dictionary interface with the implementation of the long S-fix, the majority of the f/s confusions have been eliminated.

Error types in the improved OCR

After OCR'ing with historical lexicon, noise is less coloured:

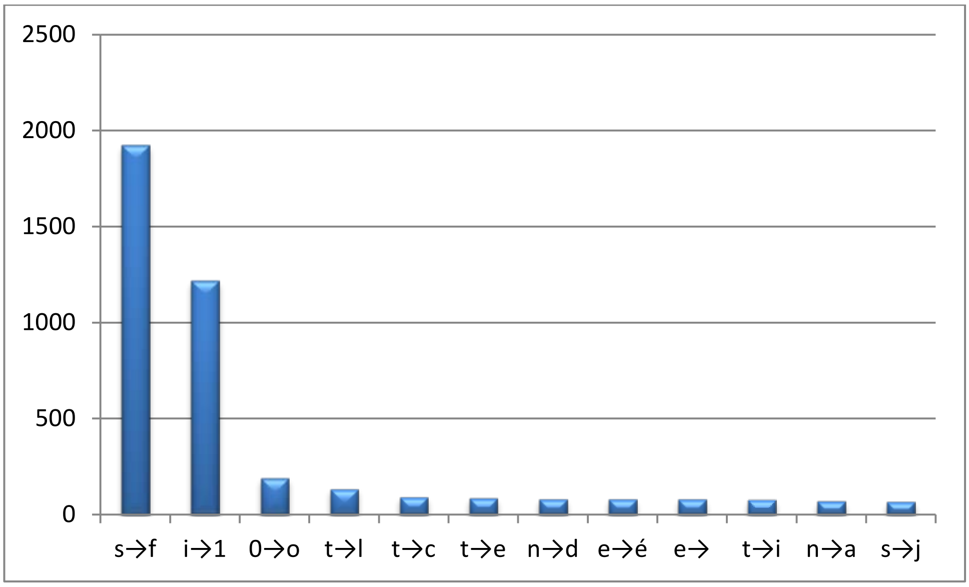


Figure 8: Frequent confusions in OCR with historical lexicon: (about 9150 confusions)

Thus, whereas the most frequent confusions contribute about 2/3 of the errors with default Finereader setup, adding the historical dictionary reduces this proportion to about 45%.

2.3.4 Results of naïve post-correction of the baseline OCR

The result of this simple procedure is as follows: Precision 0.9028, Recall 0.9063.

2.4 Discussion: summary of results

The recognition of ´long s' as f is the most frequently complained-about error in OCR of Dutch historical documents. As can be seen from the above, a the error rate can be reduced significantly by using the historical lexicon in combination with the "long s fix".

The chart below summarizes the results.

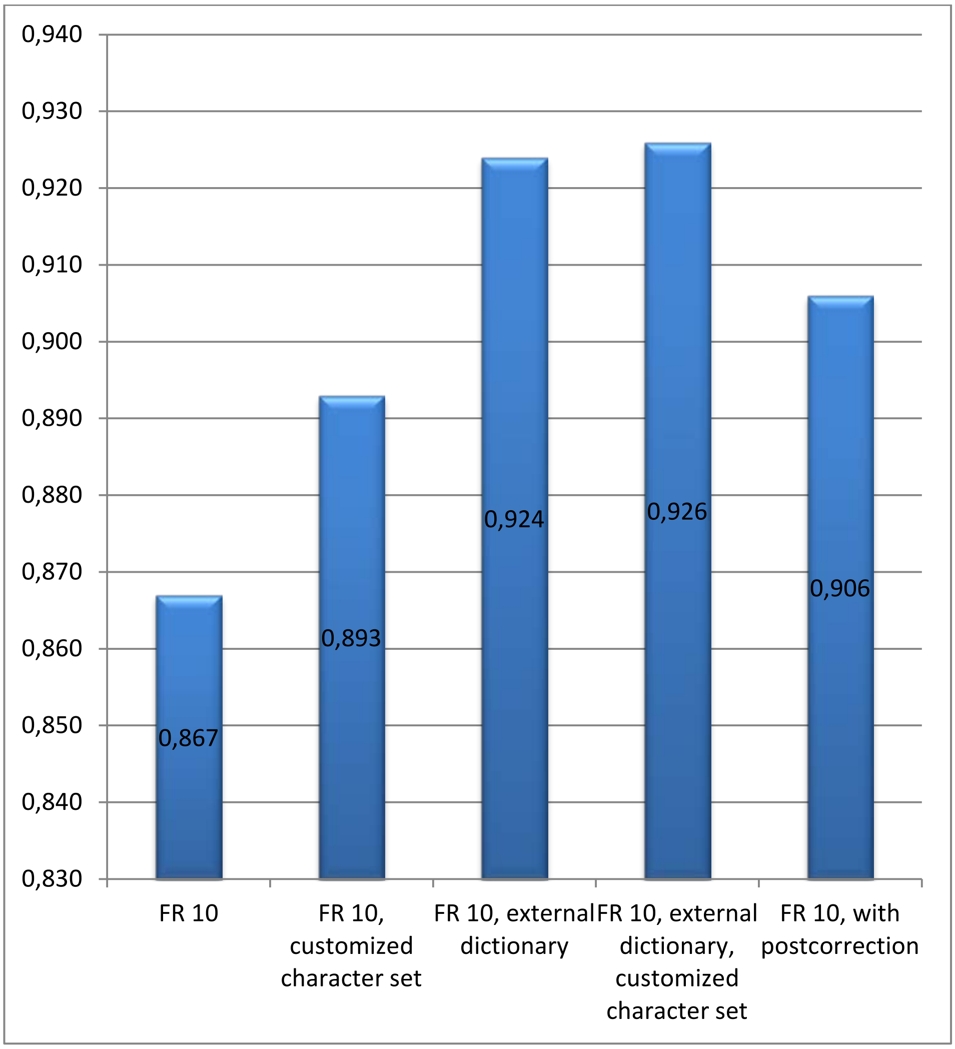


Figure 9: comparison of options for improving recognized text quality

Thus, results obtained by re-OCRing with external dictionary are best. Automatic post-correction is a good second option if this is not feasible. But one should take into account that the two are not mutually exclusive, and that it should be possible to achieve more significant progress by using a more advanced post-correction system.

It should also be noted that customization of the OCR character set is obviously worthwhile when processing a collection of historical documents. It is striking that even uptake of such simple options is an issue which requires attention.

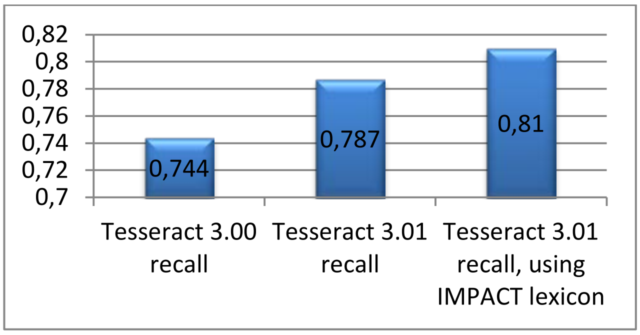
3. FUTURE WORK

IMPACT aimed to improve OCR of historical texts. Significant progress has been made but there is still a lot more work to be done.

An important task is to disseminate the results and make them available for both the research community and mass digitization. Both the lexical data on which this paper depends and the software (the module to be used with FineReader SDK) will be made available by the IMPACT centre of competence, http://www.digitisation.eu.

Obviously, another task is to extend the results of the language work in IMPACT to other OCR engines. First candidate for this would be Tesseract. Unfortunately it has not yet been feasible for us to check the influence of lexical data on Tesseract performance as thoroughly as we would like; it is not possible implement a fix for the long s problem without either retraining the engine, or modifying the implementation of lexical support in Tesseract, both of which options are outside the scope of this study. We have only been able to perform a first test by replacing the default word list in the Dutch language data.

Tesseract performance (word recall) on the Dutch Pilot evaluation set set



References

[1] Choudhury, S., Dilauro, T., Ferguson, R., Droettboom, M. and Fuginaga, I., “Document recognition for a million books”, D-Lib Magazine, Vol. 12(3) (2006).
[2] Clausner, C., Pletschacher, S. and Antonacopoulos, A., “Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments”, Proc. ICDAR 2011, 48-52 (2011).
[3] Craig, H. and Whipp, R. (2010), “Old spellings, new methods: automated procedures for indeterminate linguistic data”, Lit. Linguist Computing 25(1), 37-52 (2010) .
[4] Ernst-Gerlach, A. and N. Fuhr, N, “Generating search term variants for text collections with historic spellings”, Proc. 28th European Conference on Information Retrieval Research (ECIR 2006), (2006).
[5] Ernst-Gerlach, A. and N. Fuhr, N, “Retrieval in text collections with historic spelling using linguistic and spelling variants”, Proc. JCDL ’07, 333–341 (2007).
[6] Erjavec, T., Ringlstetter, C., Žorga, M. and Gotscharek, A., "Towards a Lexicon of XIXth Century Slovene". Proc. IS-JT '10 conference, (2010).

[7] Erjavec, T., Ringlstetter, C., Žorga, M. and Gotscharek, A., "A lexicon for processing archaic language: the case of XIXth century Slovene", Proc. WoLeR 2011 at ESSLLI, International Workshop on Lexical Resources, (2011).
[8] Erjavec, T., "Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene.", Proc. 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities, 33-38 (2011).
[9] Federico, M., Bertoldi, N and Cettolo, M., "IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models", Proc. Interspeech, Brisbane, Australia, (2008). (http://www.keithv.com/software/srilm/)
[10] Furrer, L. and Volk, M., "Reducing OCR Errors in Gothic-Script Documents", Proc. RANLP 2011 workshop on Language Technologies for Digital Humanities and Cultural Heritage, 97-103 (2011).
[11] Gotscharek, A., Neumann, A., Reffle, U., Ringlstetter, C. and Schulz, K.., "Enabling Information Retrieval on Historical Document Collections - the Role of Matching Procedures and Special Lexica", Proc. ACM SIGIR 2009 Workshop on Analytics for Noisy Unstructured Text Data (AND 2009) (2009).
[12] Gotscharek, A., Neumann, A., Reffle, U., Ringlstetter, C. and Schulz, K.., "Constructing a Lexicon from a Historical Corpus", Proc. Conference of the American Association for Corpus Linguistics (AACL09), (2009).
[13] Gotscharek, A., Neumann, A., Reffle, U., Ringlstetter, C. and Schulz, K., "On Lexical Resources for Digitization of Historical Documents", Proc. 9th ACM Symposium on Document Engineering (DOCENG 2009), (2009).
[14] Gotscharek, A., Neumann, A., Reffle, U., Ringlstetter, C., Schulz, K. and Neumann, A., "Towards information retrieval on historical document collections: the role of matching procedures and special lexica", International Journal on Document Analysis and Recognition 14, 159-171 (2010).
[15] Gravenhorst, C.: "Applied IMPACT - Does the new FineReader Engine and Dutch lexicon increase OCR accuracy and production efficiency? A case study by KB and CCS", Final IMPACT Conference, London, 2011-10-24 (http://vimeo.com/31999737).
[16] Guenthner, F., "Electronic Lexica and Corpora Research at CIS", International Journal of Corpus Linguistics, 1(2) (1996).
[17] Kempen, S., Luther, W. and Pilz, T.. "Comparison of distance measures for historical spelling variants." IFIP (International Federation for Information Processing) 217, 295–304 (2006)
[19] Koolen, M., Adriaans, F., Kamps, J. and de Rijke, M., "A cross-language approach to historic document retrieval", Proc. 28th European Conference on Information Retrieval Research (ECIR 2006), 407–419 (2006).
[20] Kopec, G., Said, M. and Popat, K., "N-Gram Language Models for Document Image Decoding", Proc. SPIE 4670, 191 (2001).
[21] Lisowski, T. "Pisownia polska. Główne fazy rozwoju (propozycja rozdziału podrecznika do nauczania tresci historycznojezykowych na studiach I stopnia", Kwartalnik Językoznawczy 3(4), (2010)
[22] Maier-Meyer, P., [Lexikon und automatische Lemmatisierung, PhD thesis], CIS, University of Munich, Munich, (1995)
[23] Mooijaart, M. (2010). "The complete history? Dutch words in four historical dictionaries", Current projects in historical lexicography. Cambridge Scholars, Newcastle, 83-98 (2010).
[24] Neudecker, C., S. Schlarb, Z. M. Dogan, P. Missier, S. Sufi, A. Williams and K. Wolstencroft. An experimental workflow development platform for historical document digitisation and analysis. Proceedings of the 2011 Workshop on Historical Document Imaging and Processing (HIP '11). ACM, New York, NY, USA, 161-168 (2011)
[25] Pilz, T., Ernst-Gerlach, A., Kempken, S., Rayson, P. and Archer, D. (2008) "The Identification of Spelling Variants in English and German Historical Texts: Manual or Automatic?" Lit Linguist Computing 23(1), 65-72 (2008)
[26] Smith, R., "Limits on the application of frequency-based language models to OCR", Proc. ICDAR 2011, 538-542 (2011)
[27] Vincent, L., "Google Book Search: Document Understanding on a Massive Scale", Proc. ICDAR 2007, (2007)