IMPACT - D-EXT2 PILOT REPORTS

Dissemination level: PU (Public)

Document history			
Version	Status	Author	Date
1.0	Final	KB: Lotte Wilms UIBK: Günter Mühlberger, Lukas Gander BSB: Doris Skaric, Mark-Oliver Fischer LMU: Thorsten Vobl, Christoph Ringlstetter INL: Jesse de Does, Katrien Depuydt PSNC: Marcin Heliński, Miłosz Kmieciak, Tomasz Parkoła USAL: Christian Clausner, Stefan Pletschacher, Apostolos Antonacopoulos	June 2012

CONTENTS

Introduction

IMPACT has tested tools in productive environments in the last half year of the projects, e.g. with pilots of the CONCERT tools at the KB, BSB and the BL. Within the extension we have encouraged the uptake of mature tools and resources by running a number of pilots, involving more tools as well as more content holders. Each pilot has been described extensively in a pilot report.

Pilots

Functional Extension Parser – UIBK (DNB)

UIBK carried out two pilots during the extension of the IMPACT project. The two main objectives of these pilots were to:

- Integrate the FEP into the digitisation workflow of the eBooks on Demand EOD Network.
- Use the FEP platform for a new application area, e.g. the automated extraction of metadata from title pages of doctoral thesis for the German National Library (DNB).

Both pilots were carried out successfully and the objectives were achieved.

Error profiler and Post Correction Module – LMU (BSB, UA, KB)

In February and March of 2012 the LMU IMPACT team conducted a series of user experiments together the three IMPACT partners of Bavarian State Library (BSB), Dutch National Library (KB) and the University of Alicante (UA). The goal of these experiments was to evaluate the capabilities of the CIS postcorrection system in comparison to other correction systems, to see how much the error profiles and batch-correction features introduced by the tool can improve the correction process. The pilot was also used to gather information about the usability of the tool and possible ways of improving it.

Post correction tools – KB (LMU, IBM, INL, KB, University of Rouen)

The KB has digitised more than 2 million pages of 10.000 books from 1780 – 1800 for the project Early Dutch Books Online[1] (EDBO). Due to the disappointing OCR quality, the Project Board decided to have students manually correct as much text as possible. Next to this manual correction, a pilot was done together with IMPACT to look at the various possibilities to enhance the OCR with the help of the following tools:

- CIS Postcorrection system with error profiling (TR5)
- Re-OCR-ing with:
  - ABBYY FRE 10
  - ABBYY FRE 10 and Dutch historical lexicon
  - IBM's Adaptive OCR
- Alto Edit, a tool developed at the KB
- PlaIR platform from the University of Rouen (an improved version of the Trove newspaper tool)

The pilot also tested the CONCERT system made by IBM, but due to the setup of the pilot the test did not do justice to CONCERT and the results were not representative of the tool. Thus, the choice was made to not use the CONCERT output in this evaluation, but instead look at the results when using IBM's Adaptive OCR compared to ABBYY 10.

The purpose of the experiment was to get a better view on the various OCR correction tools that are available, the amount of effort needed from a library and the volunteers to use such tools, and the quality of the results that can be obtained.

[1] http://www.earlydutchbooksonline.nl/

ABBYY Finereader10 vs. Tesseract – PSNC

PSNC undertook a pilot to compare the OCR accuracy of two well known OCR engines: Tesseract 3.0.1 and FineReader10 Corporate Edition. The comparison is based on Polish historical printed documents and ground-truth produced within the scope of the IMPACT project.

Dewarping pilot report – USAL

This report summarises the results of the Dewarping Pilot which was carried out in the scope of the 2012 IMPACT extension. The main goals of this pilot were to investigate the potential of dewarping for image enhancement in general and as a pre-processing step in OCR workflows in particular, based on the tools developed in IMPACT. A secondary (but very important goal) was to investigate different evaluation methodologies for dewarping evaluation.

Retrieval Application - INL

INL focused on several aspects during the extension:

- Lexicon deployment in text recognition
    - Productization of external dictionary interface. For CCS GmbH, the historical Dutch OCR lexicon has been packaged with enhanced documentation and an updated version of the INL external dictionary interface implementation.
    - OCR with Named Entity lexica. A report on the contribution of the NE lexica to text recognition.
    - Using morphology in OCR. Test and evaluation of the module developed for deploying finite state morphology in OCR during the project.
- Using the INL Retrieval application developed in IMPACT, INL validated improved access. The emphasis was on scalability and deployment in real-life situations, with the addition / enhancement of several functionalities. For the specific languages, the following work was done:
    - Dutch: the Early Dutch Books Online (EDBO) collection, consisting of about 2 million pages from 10.000 books, has been indexed and deployed in the application to put the scalability to the test. This set has been the main test set for the optimization of the application.
    - Polish: pilot based on the Nowe Ateny encyclopedia, involving the use of the Polish IR lexicon.
    - Spanish: pilot based on data from the Biblioteca Virtual Miguel de Cervantes.

KB Pilot Report
IMPACT Extension

Author:	Lotte Wilms
Date:	04/07/2012
Version:	1.2
Status:	Final

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands

TABLE OF CONTENTS

Executive summary

The KB has digitised more than 2 million pages of 10.000 books from 1780 – 1800 for the project Early Dutch Books Online[1] (EDBO). Due to the disappointing OCR quality, the Project Board decided to have students manually correct as much text as possible. Next to this manual correction, a pilot was done together with IMPACT to look at the various possibilities to enhance the OCR with the help of the following tools:

- CIS Postcorrection system with error profiling
- Re-OCRing with:
  - ABBYY FRE 10 and Dutch historical lexicon
  - IBM's Adaptive OCR
- Alto Edit, a tool developed at the KB
- PlaIR platform from the University of Rouen (an improved version of the Trove newspaper tool)

The pilot also tested the CONCERT system made by IBM, but due to the setup of the pilot the test did not do justice to CONCERT and the results were not representative of the tool. Thus, the choice was made to not use the CONCERT output in this evaluation, but instead look at the results when using IBM's Adaptive OCR compared to ABBYY 10. After the pilot, IBM has chosen process all sets with the CONCERT tool without any time restraints to make a general comparison possible. This has been added in a separate paragraph.

The purpose of the experiment was to get a better view on the various OCR correction tools that are available, the amount of effort needed from a library and the volunteers to use such tools, and the quality of the results that can be obtained.

This report describes the setup and outcomes of the pilot. The goals of the pilot have been achieved. All tools tested did contribute towards a higher OCR word accuracy rate. One of the most important conclusions that can be drawn from this pilot is that it depends on the goals set by the library which tool is the best choice for an OCR correction project. Ideally it would be a combination of re-OCRing and post correction, which will provide the best possible results.

[1] http://www.earlydutchbooksonline.nl/

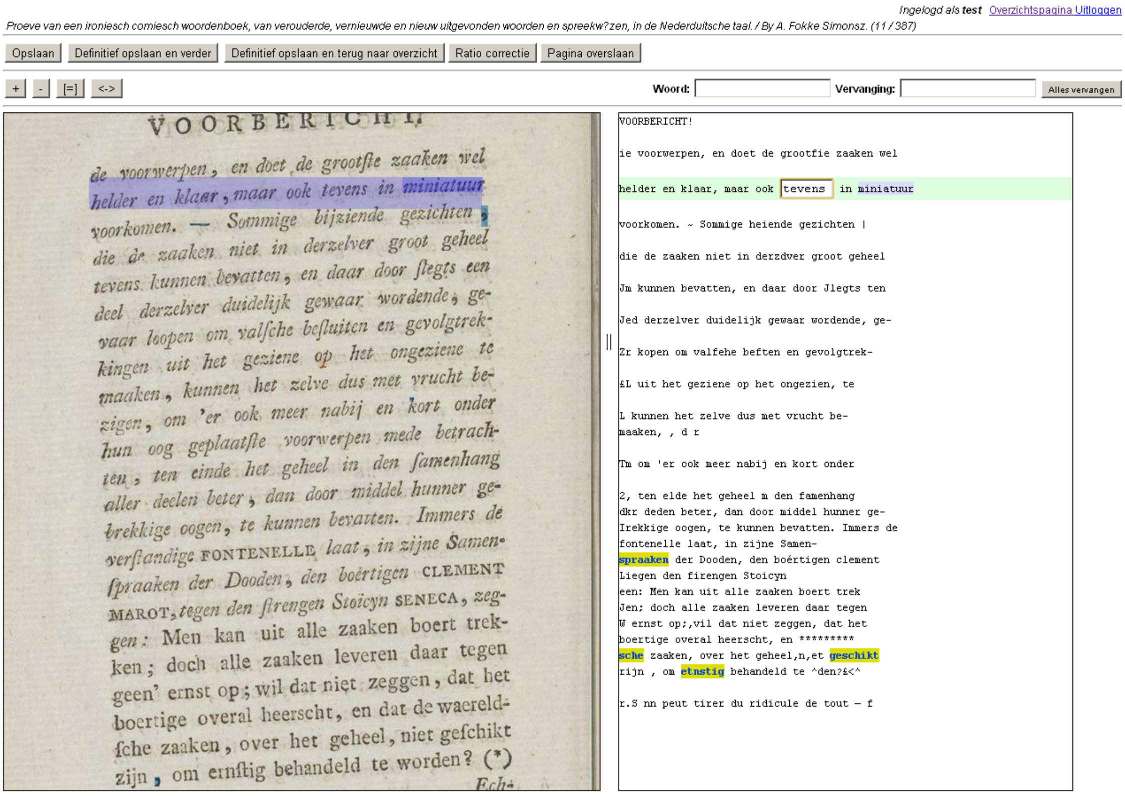
1. Tools

1.1 KB Alto Edit[2]

Within the KB, the Research department worked on improving and correcting OCR results. A part of this research involved the development of a web-based OCR correction tool; Alto Edit. The tool uses the so called side-by-side method, where the user has the scan on the left side of the screen and the OCR on the right.

The texts for this tool have been corrected on long s/f-errors, using a rule based automated correction. Texts from this period use the long s, which is often recognised as an f. On the basis of a few simple rules in the OCR, it becomes possible to replace any unlikely recognised f by an s. The users can then, where necessary, undo this replacement by correcting the word.

By clicking a wrongly recognised word in the OCR field, the word is opened in a textbox where the user can then correct it. The word is simultaneously highlighted on the scan. It is also possible to correct one word through the whole book with a search-and-replace function, seen on the top right in Figure 1. By typing a space within a word, or deleting a whole word, the user is presented with the option to correct the segmentation, as seen in Figure 2.



Proeve van een ironiesch comiesch woordenboek, van verouderde, vernieuwde en nieuw uitgevonden woorden en spreekw?zen, in de Nederduitsche taal. / By A. Fokke Simonsz. (11 / 387)

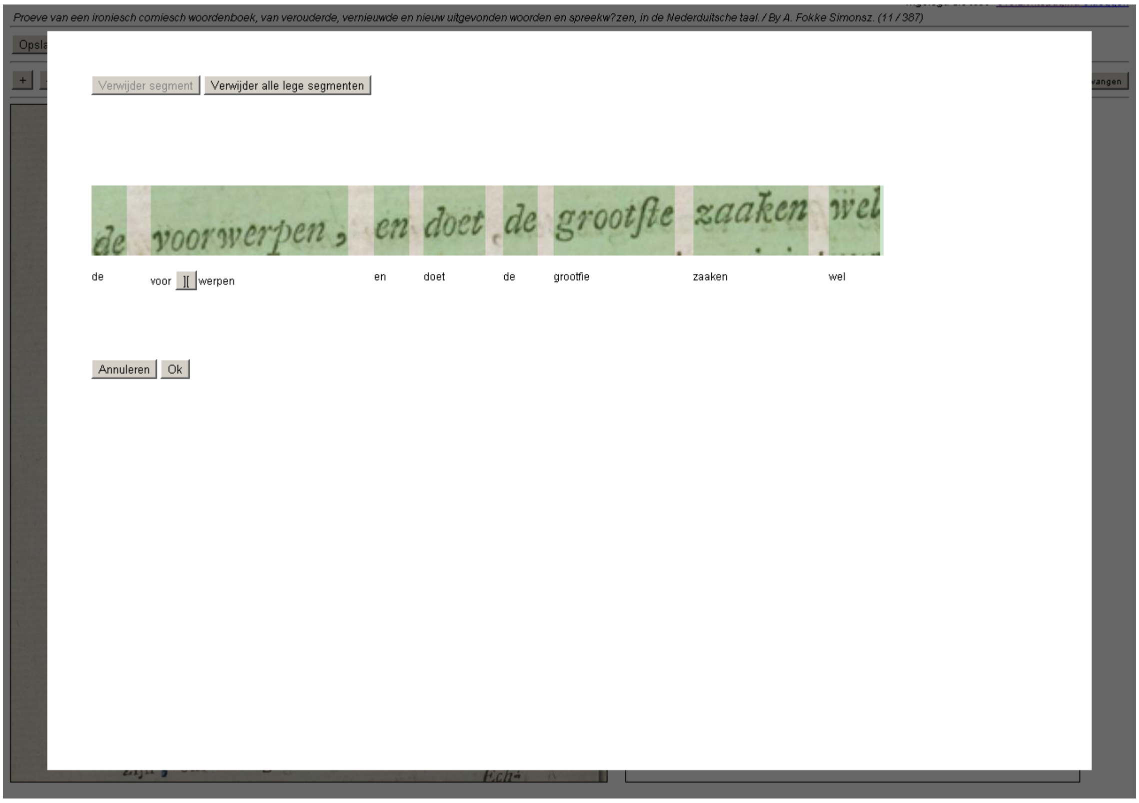
VOORBERICHT.

de voorwerpen, en doet de grootſte zaaken wel
helder en klaar, maar ook tevens in miniatuur
voorkomen. — Sommige bijziende gezichten,
die de zaaken niet in derzelver groot geheel
tevens kunnen bevatten, en daar door ſlegts een
deel derzelver duidelijk gewaar wordende, ge-
vaar loopen om valſche beſluiten en gevolgtrek-
kingen uit het geziene op het ongeziene te
maaken, kunnen het zelve dus met vrucht be-
zigen, om 'er ook meer nabij en kort onder
hun oog geplaatſte voorwerpen mede betrach-
ten, ten einde het geheel in den ſamenhang
aller deelen beter, dan door middel hunner ge-
brekkige oogen, te kunnen bevatten. Immers de
verſtandige FONTENELLE laat, in zijne Samen-
ſpraaken der Dooden, den boërtigen CLEMENT
MAROT, tegen den ſtrengen Stoïcyn SENECA, zeg-
gen: Men kan uit alle zaaken boert trek-
ken; doch alle zaaken leveren daar tegen
geen' ernst op; wil dat niet zeggen, dat het
boertige overal heerscht, en dat de waereld-
ſche zaaken, over het geheel, niet geſchikt
zijn, om ernſtig behandeld te worden? (*)
Ech-

```
VOORBERICHT!

ie voorwerpen, en doet de grootfie zaaken wel

helder en klaar, maar ook tevens in miniatuur

voorkomen. ~ Sommige heiende gezichten |

die de zaaken niet in derzdver groot geheel

Jm kunnen bevatten, en daar door Jlegts ten

Jed derzelver duidelijk gewaar wordende, ge-

Zr kopen om valfehe beften en gevolgtrek-

£L uit het geziene op het ongezien, te

L kunnen het zelve dus met vrucht be-
maaken, , d r

Tm om 'er ook meer nabij en kort onder

2, ten elde het geheel m den famenhang
dkr deden beter, dan door middel hunner ge-
Irekkige oogen, te kunnen bevatten. Immers de
fontenelle laat, in zijne Samen-
spraaken der Dooden, den boértigen clement
Liegen den firengen Stoicyn
een: Men kan uit alle zaaken boert trek
Jen; doch alle zaaken leveren daar tegen
W ernst op;,vil dat niet zeggen, dat het
boertige overal heerscht, en *********
sche zaaken, over het geheel,n,et geschikt
rijn , om etnstig behandeld te ^den?£<^

r.S nn peut tirer du ridicule de tout — f
```

Figure 1 – KB Alto Edit main screen

[2] https://github.com/impactcentre/alto-editor



Proeve van een ironiesch comiesch woordenboek, van verouderde, vernieuwde en nieuw uitgevonden woorden en spreekw?zen, in de Nederduitsche taal. / By A. Fokke Simonsz. (11 / 387)

Figure 2 – Segmentation correction in Alto Edit

1.2 LMU Profiler and Post Correction Tool

During the IMPACT Project (2008-2011), a working group at the University of Munich developed software to analyse the OCR output from historical documents, using statistical modelling of document characteristics to improve OCR accuracy.

The statistical models are as follows:

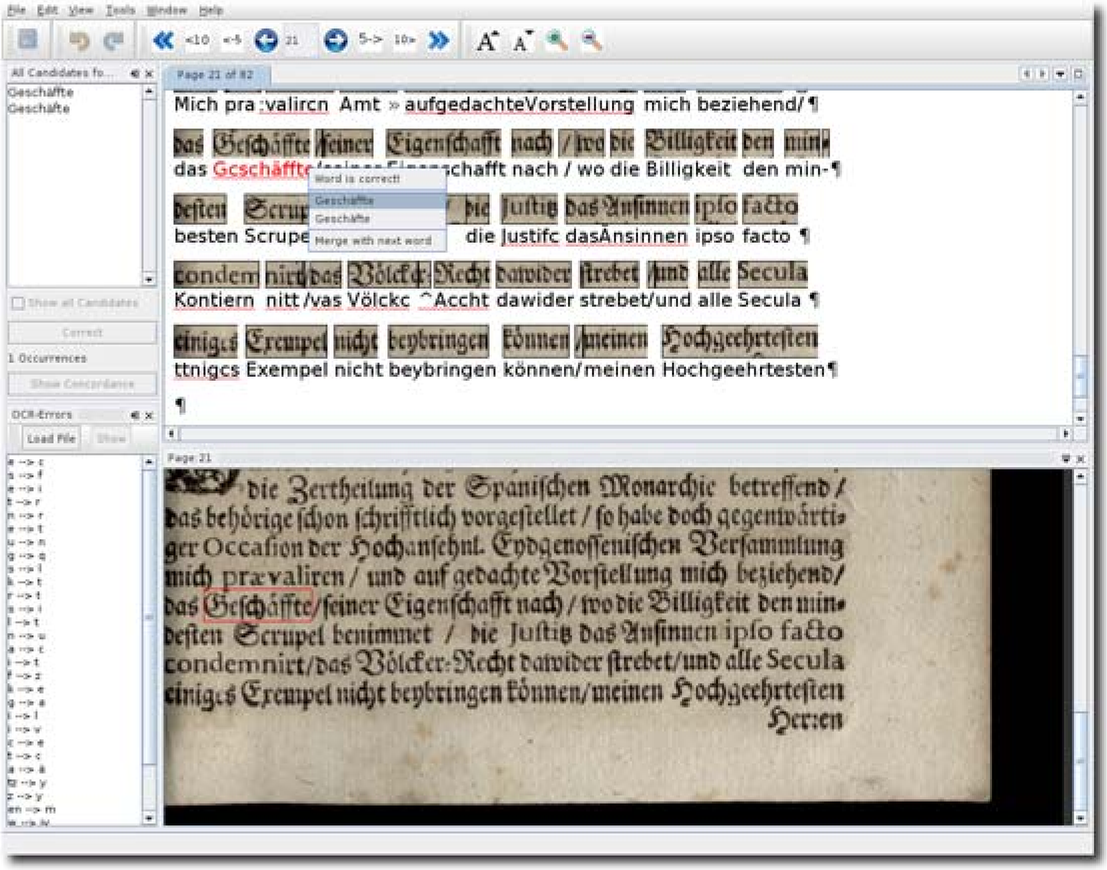
1. An analysis of the vocabulary of the document, focussing on potential variant spellings, and on the appearance of secondary or tertiary languages in the text.
2. An analysis of known language rules or patterns that can explain variant spellings.
3. An analysis of the OCR error rate, focussing on systematic errors (ie, those that appear to be introduced by the process of OCR itself).
4. Related to this, an analysis of which error patterns occur with high probability (ie, ‘i’ for ‘l’, ‘in’ for ‘m’, ‘n’ for ‘u’).

This analysis can be used for quality control, post-correction, retrieval, and a second run of the OCR that will adapt to the results produced by the Text and Error Profiler.[3]

The Text and Error Profiler works by attuning itself to a particular document, rather than to common traits of printed documents from a certain era, resulting in a highly adaptive process. In particular, the tool uses its document-specific knowledge to allow the batch processing of erroneous words. By statistical analysis of the whole document, the

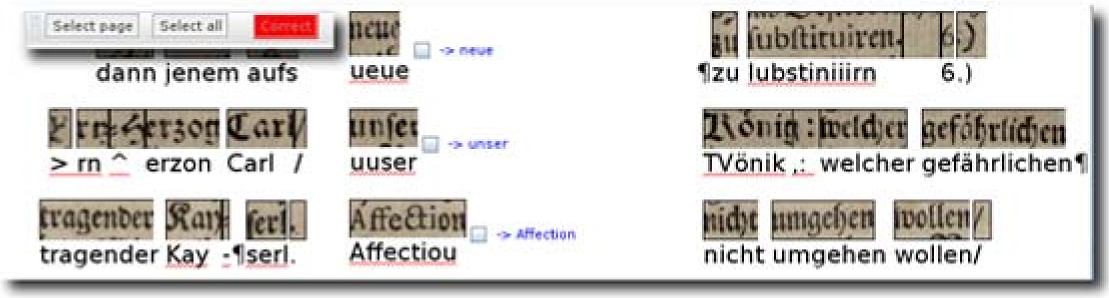
[3] http://www.digitisation.eu/tools/ocr-post-correction-and-enrichment/text-and-error-profiler/

system can identify correct substitute words with high confidence and huge numbers of systematic errors can be post-corrected with just a few keystrokes.[4]



Mich pra :valircn Amt » aufgedachteVorstellung mich beziehend/ ¶
das Gcschäffte /... ...chafft nach / wo die Billigkeit den min-¶
besten Scrupe... die Justifc dasÄnsinnen ipso facto ¶
Kontiern nitt /vas Völckc ^Accht dawider strebet/und alle Secula ¶
ttnigcs Exempel nicht beybringen können/meinen Hochgeehrtesten¶
¶

Figure 3 – Overview LMU Post Correction Tool



dann jenem aufs
ueue
¶zu lubstiniiirn 6.)
> rn ^ erzon Carl /
uuser
TVönik ,: welcher gefährlichen¶
tragender Kay -¶serl.
Affectiou
nicht umgehen wollen/

Figure 4 – Batch correction by error pattern (n>u)

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

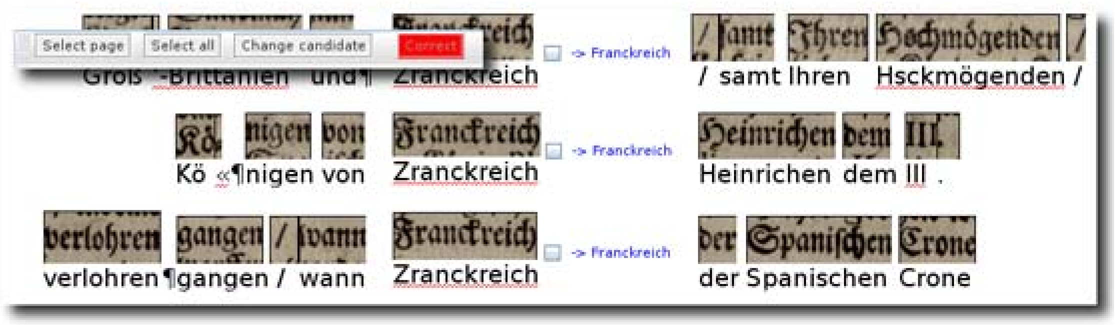


Figure 5 – Batch correct by word

1.3a CONCERT

The COoperative eNgine for Correction of ExtRacted Text (CONCERT) is a web-based platform, suitable for massive volunteer participation, which validates and corrects OCR results. In this way, it enables the general public to help with large scale digitisation efforts.

The technology streamlines, simplifies and accelerates the process of winnowing out questionable text scans, enabling reviewers to key in corrections to the text. Instead of displaying an entire scanned page, reviewers only see the actual letters or words in question. For example, the letter combination "r" and "n" ("rn") may appear indistinguishable from the letter "m." In those instances, the system collects many instances of the letter "m," and places these samples next to the letters in question, making it much easier to determine the letter's real identity (see Figure 7).

In cases where an entire word is suspect, it is added to a collection of other questionable terms, which are then arranged in alphabetical order (see Figure 8). Volunteer reviewers need only accept or reject suggested substitutes with one keystroke. In addition, the system uses adaptive dictionary enrichment, a method in which new words are added to a central dictionary based on cross-identification and correction by other users.

In the final session of the tool, operators are shown the full page in context (see Figure 9). This will allow them to correct any outstanding letters, to identify false positives, and to correct the segmentation of the page (where words and letters have been incorrectly combined or dispersed).[5]

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

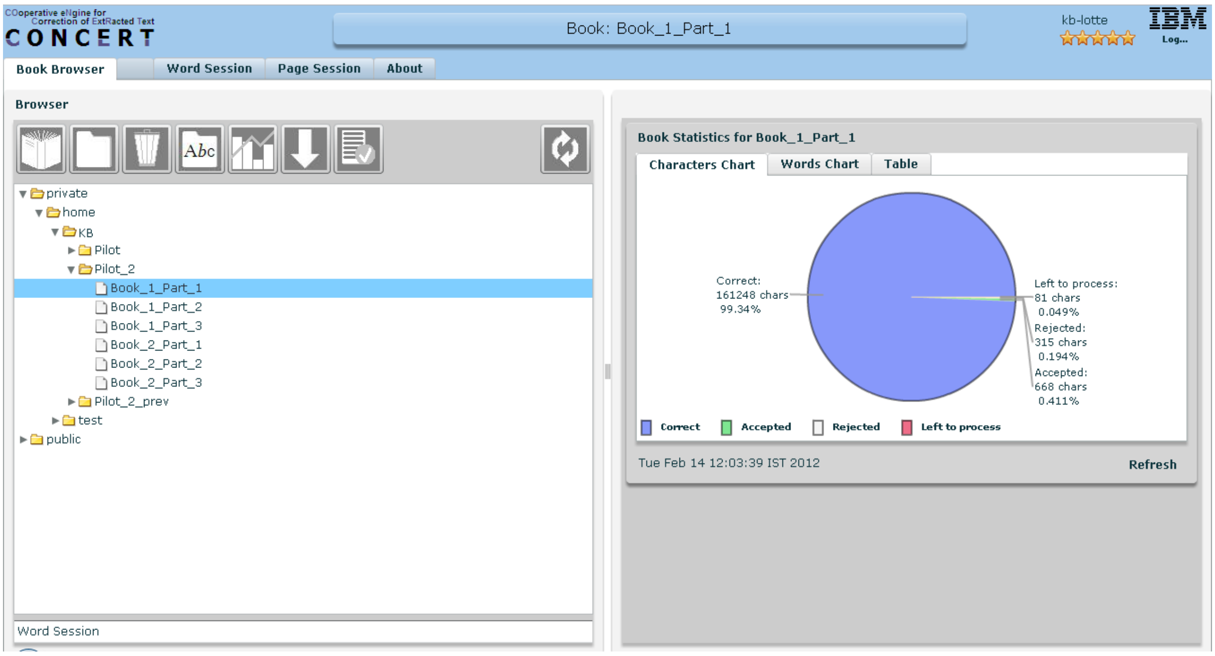


Figure 6 – Overview CONCERT

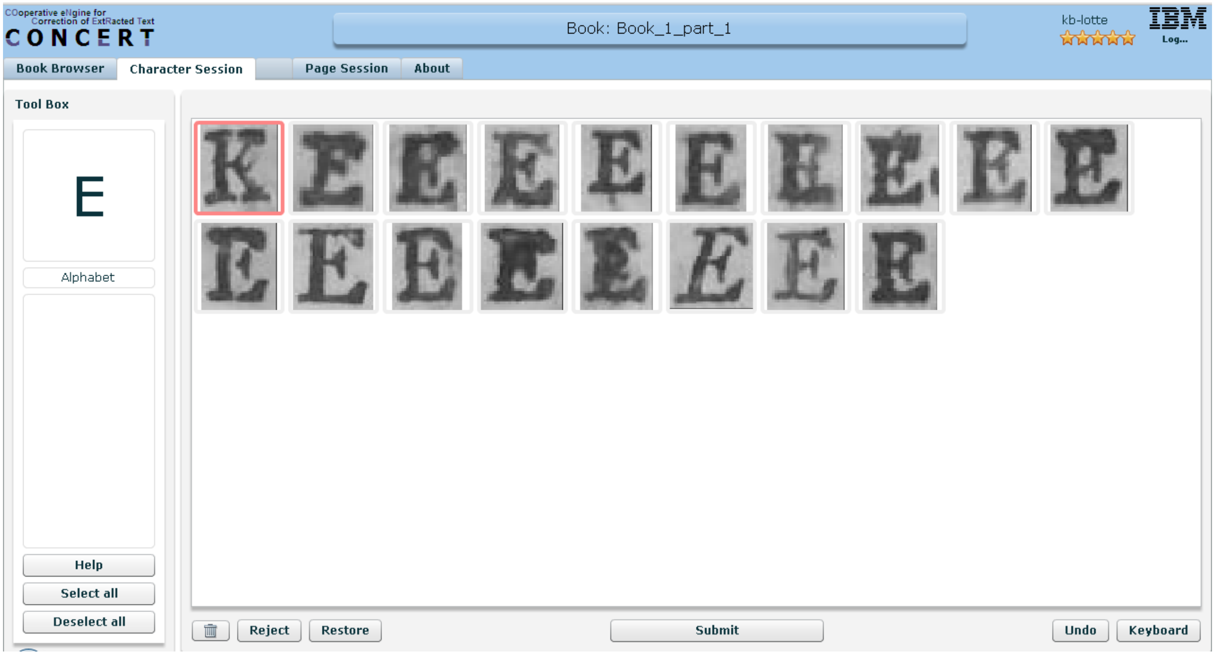


Figure 7 – Character session

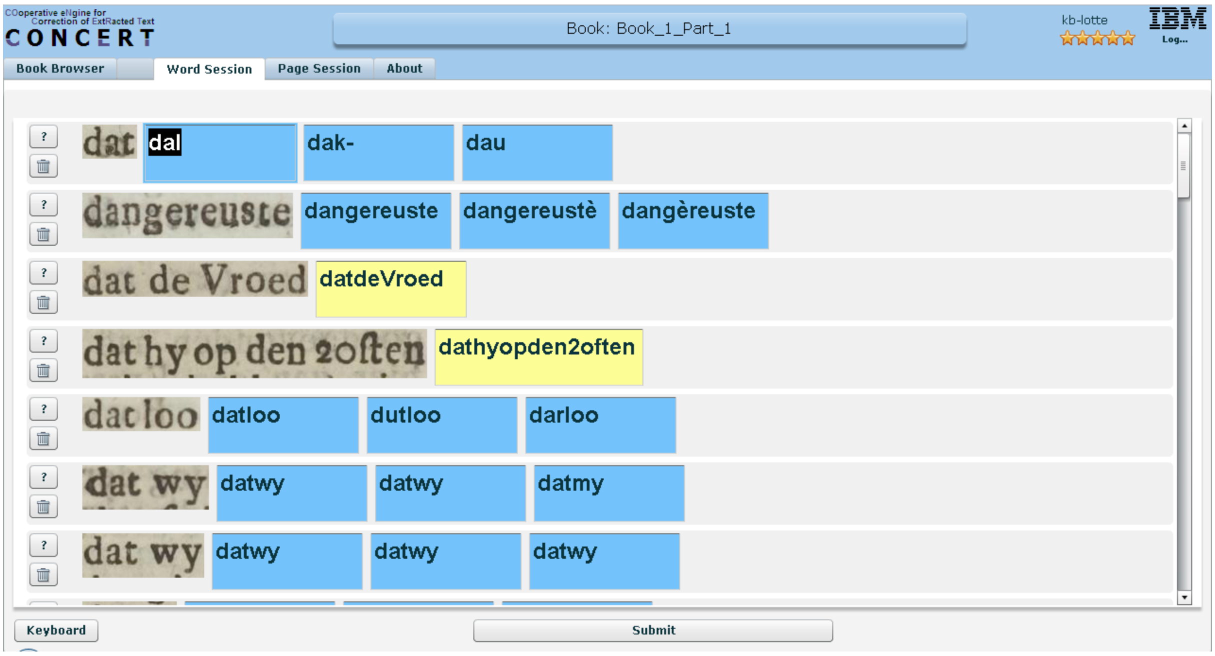


Figure 8 – Word session

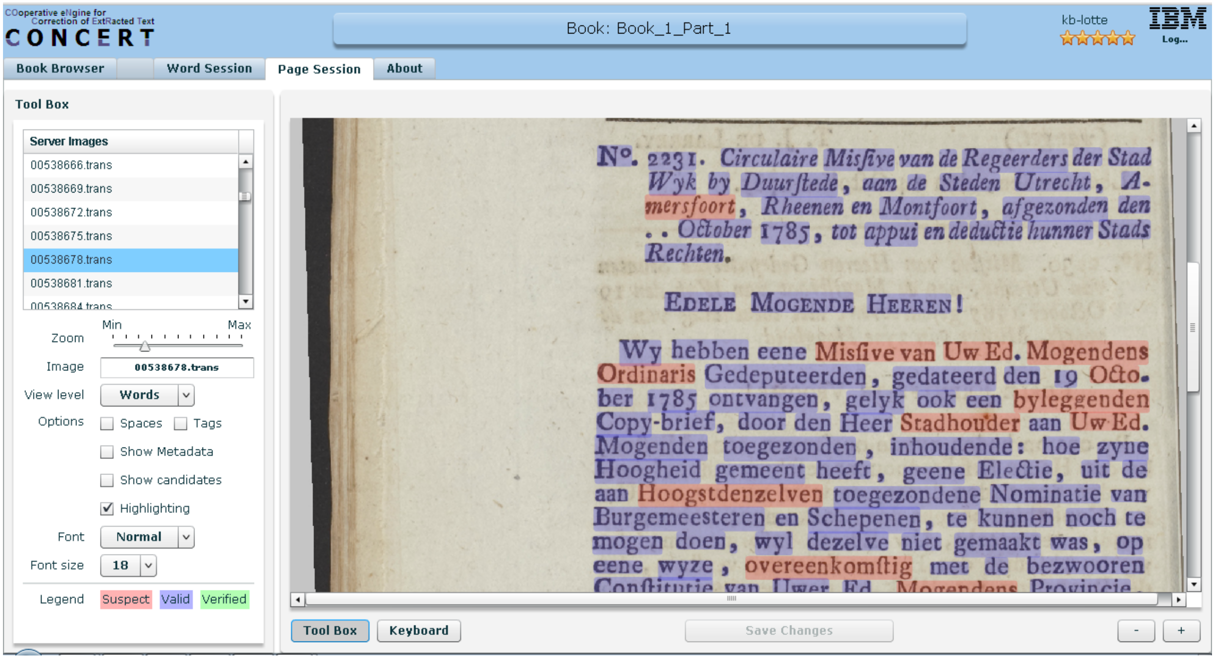


Image: 00538678.trans
View level: Words
Options: □ Spaces □ Tags
□ Show Metadata
□ Show candidates
☑ Highlighting
Font: Normal
Font size: 18

N°. 2231. Circulaire Misſive van de Regeerders der Stad Wyk by Duurſtede, aan de Steden Utrecht, Amersfoort, Rheenen en Montfoort, afgezonden den . . October 1785, tot appui en deductie hunner Stads Rechten.

EDELE MOGENDE HEEREN!

Wy hebben eene Misſive van Uw Ed. Mogendens Ordinaris Gedeputeerden, gedateerd den 19 October 1785 ontvangen, gelyk ook een byleggenden Copy-brief, door den Heer Stadhouder aan Uw Ed. Mogenden toegezonden, inhoudende: hoe zyne Hoogheid gemeent heeft, geene Electie, uit de aan Hoogstdenzelven toegezondene Nominatie van Burgemeesteren en Schepenen, te kunnen noch te mogen doen, wyl dezelve niet gemaakt was, op eene wyze, overeenkomſtig met de bezwooren Conſtitutie van Uwer Ed. Mogendens Provincie.

Figure 9 – Page session

Due to the setup of the pilot, this tool was not tested in the way it was intended to be used, resulting in misrepresentative outcomes. CONCERT is developed to work with input from a large number of volunteers on a large selection of material. The processing time between sessions can take some time, which means a certain set is not available. When working on multiple sets with multiple people, this is not a problem, as a volunteer can simply work on another set. However, with the setup of the pilot, this was not possible. During the pilot it indeed became a *problem that some sets were not available for a longer period of time. The correction of the OCR thus had to be stopped at a crucial stage in the process of CONCERT, making the output incomparable to other correction methods. However, a comparison could be made between the Adaptive OCR software developed by IBM and ABBYY 10 (see paragraph 4.1.4 and 4.2.4). After the pilot, IBM has chosen to process the books completely using CONCERT, thus making a general comparison possible (see paragraph 4.4).*

1.3b Re-OCRing with IBM's Adaptive OCR

IBM Adaptive OCR is a comprehensive software system which improves the recognition of historical texts significantly by applying adaptivity as one of the main features to the text recognition process.
It integrates several other tools, such as the image enhancement toolkit, the ABBYY FineReader Engine, the post correction tool and the lexical resources developed during the IMPACT project.[6]

1.4 PlaIR (LITIS Laboratory - Rouen)

The PlaIR project: A platform for the digitisation, organisation and retrieval of old newspaper archives

The *Journal de Rouen* is one of the oldest corpora among the old regional newspapers in France. It covers the years 1762-1947 and has been digitized by the Upper Normandy Archives. Digitization was performed so as to provide high-resolution images (300.000 images for the whole corpus at a resolution of 300 DPI) as well as encapsulated PDF files that include both compressed images and OCR results for each daily issue. On line access to this digitized corpus is limited due to the following reasons:

- Only indexation by date is possible.
- Textual queries can only be performed one issue at a time by means of the encapsulated PDF files.
- Indexing is prone to many errors due the low performance of the OCR used during the digitization process.

LITIS Laboratory has launched a research program supported by the Upper Normandy Council so as to provide improved indexing facilities of this corpus and web access with interactive tools.

The strength of the proposed system is to provide retrieval facilities of newspaper articles by textual queries as opposed to most of the other systems which are limited to page or issue indexing. It is composed of two major components:

- A web platform dedicated to user interaction, retrieval facilities, high-resolution visualisation of the digitized documents, and crowd sourcing allowing OCR correction.
- A digitization workflow dedicated to document image analysis and recognition featuring intelligent page segmentation into articles, OCR integration, formatted outputs in METS/ALTO files.

The online consultation platform is the front-end of our system. This browsing and visualization component relies on light web technologies. Therefore it does not require any additional plug-in neither any system parameter setting. The user has simply to use his favourite web browser to access the documents and search for the information he needs. This web platform has been designed to provide:

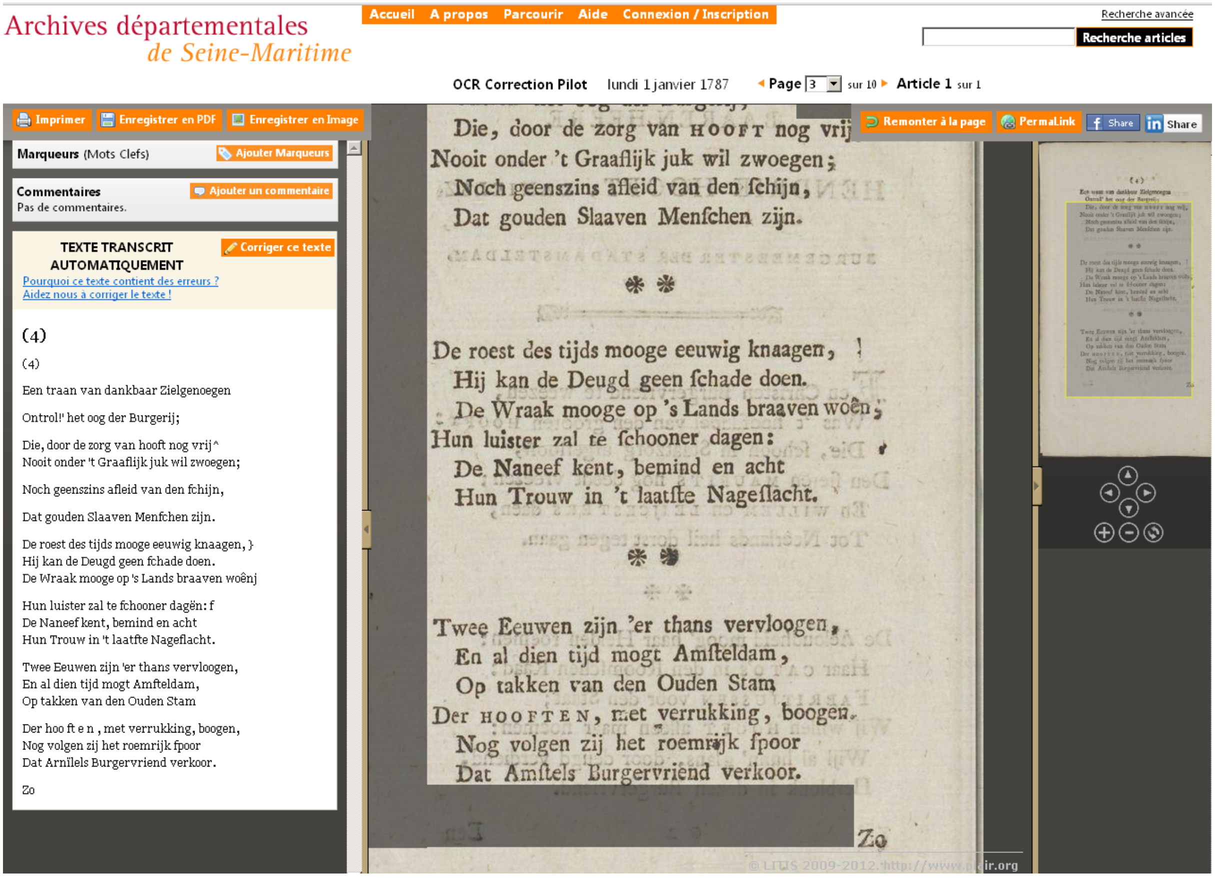
- High-resolution visualization facilities of the images with zooming and highlighting of articles
- Indexation by articles using OCR results

[6] http://www.digitisation.eu/tools/ocr-engines/ibm-adaptive-ocr-engine/

- Search facilities in the articles database by keywords with image highlighting of the retrieved results
- Collaborative OCR correction for improving the whole performance of the system

This platform is built upon Open Source technologies: IIPImage, Apache Lucene, MySQL, National Library of Australia.[7],[8]

During this pilot the option for OCR correction was used. The PlaIR platform is included in the pilot due to the good results that were presented at the IMPACT workshop, held on 31 March 2011 in Rouen. The tool was adjusted slightly to make working with books somewhat easier. This work was carried out by the University of Rouen.



Pas de commentaires.

TEXTE TRANSCRIT AUTOMATIQUEMENT

(4)

(4)

Een traan van dankbaar Zielgenoegen

Ontrol!' het oog der Burgerij;

Die, door de zorg van hooft nog vrij^
Nooit onder 't Graaflijk juk wil zwoegen;

Noch geenszins afleid van den fchijn,

Dat gouden Slaaven Menfchen zijn.

De roest des tijds mooge eeuwig knaagen, }
Hij kan de Deugd geen fchade doen.
De Wraak mooge op 's Lands braaven woênj

Hun luister zal te fchooner dagën: f
De Naneef kent, bemind en acht
Hun Trouw in 't laatfte Nageflacht.

Twee Eeuwen zijn 'er thans vervloogen,
En al dien tijd mogt Amfteldam,
Op takken van den Ouden Stam

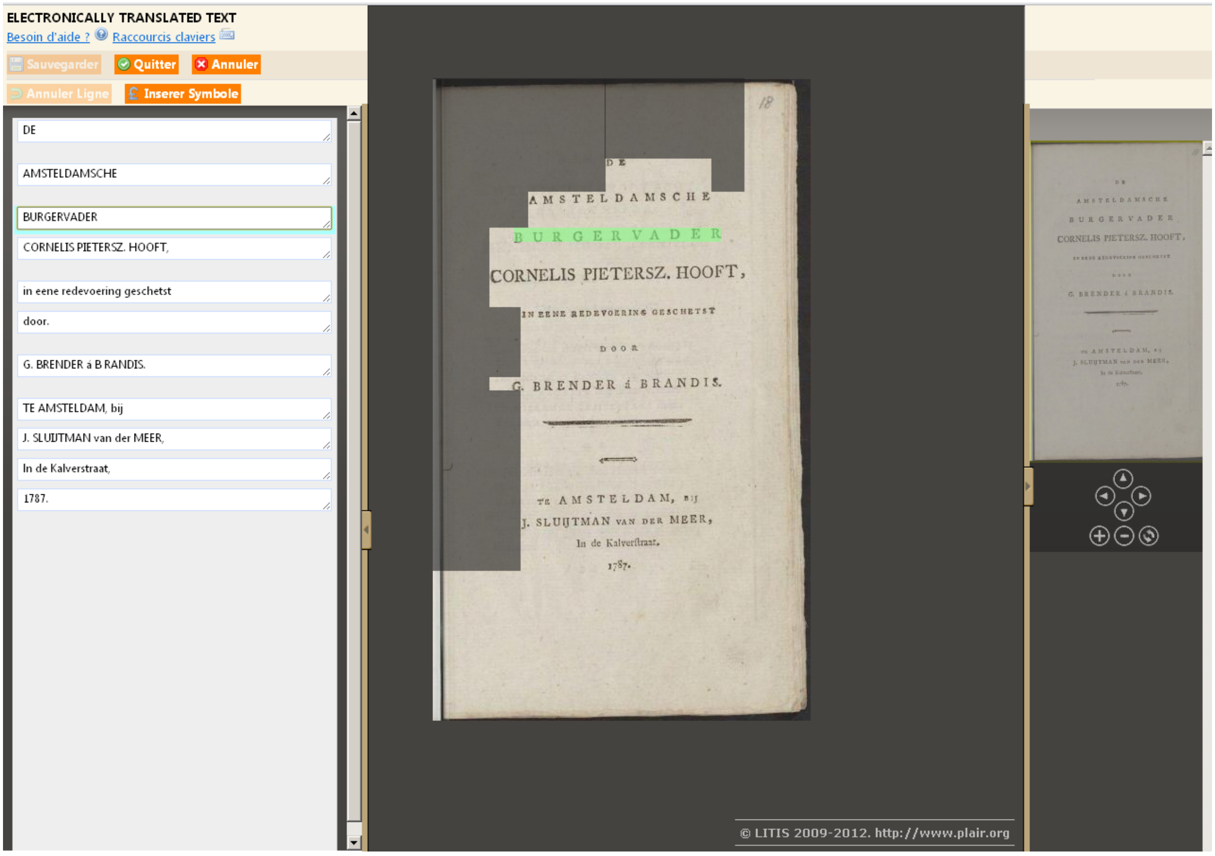
Der hoo ft e n , met verrukking, boogen,
Nog volgen zij het roemrijk fpoor
Dat Arnîlels Burgervriend verkoor.

Zo

Figure 10 – Overview PlaIR platform

[7] http://plair.univ-rouen.fr/plair/jdr/home

[8] http://www.plair.org



DE

AMSTELDAMSCHE

BURGERVADER

CORNELIS PIETERSZ. HOOFT,

in eene redevoering geschetst

door.

G. BRENDER a B RANDIS.

TE AMSTELDAM, bij

J. SLUIJTMAN van der MEER,

In de Kalverstraat,

1787.

Figure 11 – Correction mode

1.5 Re-OCRing with ABBYY FRE 10 and IMPACT Historical Dictionary[9]

Through IMPACT, the state-of-the-art OCR engine ABBYY FineReader has been adapted to cope with the challenge of recognising historical fonts and layouts.

The new SDK FineReader Engine 10, which was released in September 2010, contains a variety of technological improvements in terms of processing speed, recognition accuracy, simplification of development and new export formats. New intelligent binarisation of the document images makes sure that more text is transferred to the OCR process. The new technology was also developed and tested with historical books provided by the other IMPACT partners.

Improved functionalities of the ABBYY FineReader Engine include:

- A new adaptive binarisation, which works better for documents with non-uniformly coloured background, noise and bleed-through from the opposite pages.
- Improvements in the segmentation algorithms, for example regarding picture detection. The segmentation of historic newspapers is also improved.

[9] A new version of the implementation became available during this pilot. More information on this can be found in the Extension Report by INL. When this implementation proves to be more successful than the current version running in the framework, it will be updated.

- Improvements in Fraktur recognition, using ground truth data made available during the IMPACT project.
- Improvements in the External Dictionary interface, which now allows using user-developed dictionaries to better recognise languages not supported by ABBYY's technology.
- Native ALTO export format.[10]

Dictionaries

The various language institutes in IMPACT have worked on building lexica for historical languages. The aim has been to improve OCR results for historical text, and also to ensure that the user finds historic variants of word when searching for the modern-day form.

A lexicon is a structured, machine-usable repository of relevant linguistic knowledge about words in a language. A lexicon will contain historical variants (orthographical variants, inflected forms) and link them to a corresponding dictionary form in modern spelling (known as a 'modern lemma'). In this way, a user can search for a modern word ('water') and receive results that take into account all historical variants in that language ('wæter', 'weter', 'waterr', 'watre', etc.)[11]

These dictionaries can be used in combination with ABBYY Finereader Engine 10 to better tune the OCR software to the historical text. The lexicon used for this consists of a checked list of words, based on a corpus of dated texts, preferably also with frequency information and from the same time period or of the same text type as the text that needs to be OCRed.[12]

[10] http://www.digitisation.eu/tools/ocr-engines/abbyy-finereader-engine/

[11] http://www.digitisation.eu/tools/language-resources/

[12] http://www.slideshare.net/impactproject/katrien-depuydt?ref=http://www.digitisation.eu/tools/language-resources/

2. Material

2.1 About Early Dutch Books Online[13]

Early Dutch Books Online gives full-text access to more than 2 million pages in 10,000 books from the Dutch-speaking region from the period 1781-1800.

The project is a collaboration between the National Library of the Netherlands and the university libraries of Amsterdam and Leiden. Books from the Special Collections of these libraries have been digitized and made available on word level via th[e] website.

Online library

The Special Collections departments of UB UVA, UBL and the KB launched the initiative "National Infrastructure for Digital Access to Special Collections" in October 2005. This is a plan for an online library for Humanities consisting of fully digitized items from the Special Collections of the institutions involved. Digitizing the various Special Collections from these three libraries, and in time also from other libraries, makes a large quantity of previously mostly inaccessible texts accessible to scholars and for education. Early Dutch Books Online is the first step toward this online library.

The importance of digitization of scientific sources is evident. Without source material, research in the Humanities is impossible. Electronic access contributes to the efficiency, effectiveness and reliability of the research and provides opportunities for entirely new types of research. Digitization makes new scientific breakthroughs possible. The availability of large text corpora is necessary for this. Early Dutch Books Online makes such large files accessible.

2.2 Material selected

Two works were selected from the Early Dutch Books Online (EDBO) set for this 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
   - http://www.earlydutchbooksonline.nl/nl/view/image/id/dpo:3077:mpeg21
   - Year: 1791
   - Printer/publisher: Chalmot, Jacques Alexandre de Kampen, 1778-1797
   - Copy: Leiden, Universiteitsbibliotheek: 1006 A 29
   - No. of pages: 338

2. Verhandelingen van het Genootschap ter bevordering der heelkunde, te Amsterdam. : Part 1
   - http://www.earlydutchbooksonline.nl/nl/view/image/id/dpo:3423:mpeg21
   - Year: 1791
   - Printer/publisher: Elwe, Jan Barend Amsterdam, 1778-1800
   - Copy: Leiden, Universiteitsbibliotheek: 1448 G 2
   - No. of pages: 331

[13] http://www.earlydutchbooksonline.nl/en/edbo/page/project

The choice of books was limited due to the availability of ground-truth (99,95% correct text) produced within IMPACT. This ground-truth is needed for the evaluation of the outcomes. The two works that were selected have been chosen due to the large numbers of pages they contain.

Because of their deviant size, the foldouts from book 2 have been excluded from the pilot. Empty pages and book covers have also been excluded, due to the limitations of the evaluation software.

3. Setup

The pilot was performed with six testers, all Humanities students. The two books have been divided into six sets to make sure the testers did not became too familiar with the material. The students have each corrected a complete book, plus one set of the other book, but used different tools for each set. A set consist of a cross-section of the whole book, with pages 1-4-7 in Set 1, 2-5-8 in Set 2 and 3-6-9 in Set 3.

Unfortunately one person could not join the final test, so the decision was made to test the PlaIR tool with only five sets.

	Alto Edit (6-7 February)	LMU (9-10 February)	PlaIR (15-16 May)
Person 1	Set 1-1	Set 1-2	Set 2-1
Person 2	Set 1-2	Set 1-3	Set 2-2
Person 3	Set 1-3	Set 1-1	
Person 4	Set 2-1	Set 2-2	Set 1-1
Person 5	Set 2-2	Set 2-3	Set 1-2
Person 6	Set 2-3	Set 2-1	Set 1-3

For each tool the testers received a training, with subsequently the possibility to work with the tool and ask questions. These sessions were always done on the day before the actual test and lasted 2,5 hours. The material used for the training was not the same as used in the test, but did come from the EDBO collection. The first training session was longer and included an explanation to the whole pilot and the guidelines for the test. These guidelines are unfortunately only available in Dutch, but the most important instructions were to key the actual text of the book, including errors and historical spelling, and to separate the ligatures.

The test was carried out on the morning following the training, for which a time limit of 3 hours was set, with a 15-minute break after 1,5 hours. The test was timed using two stopwatches.

The user experience was out of scope for this pilot. Some tool providers did request the users to fill out an evaluation form, but this was only intended for their own use.

3.1 Comments regarding the setup

A cross-comparison of the different tools is rather difficult to set up – each of the tools listed is tailored to a particular use case and it is hardly possible to design the experiment in such a way that the testing conditions are equally fair for each approach, while at the same time obtaining results in such a form that they can be evaluated straightforward and using the same metrics.

Consequently, some comments have to be made with regard to the setup of the pilot:

- Everyone is (somewhat) new to the material, the guidelines and the OCR correcting. This could have adverse consequences for the first tool. To counterbalance this effect, it was decided to test the 'easiest' tool first, which required the least amount of training to give the testers the opportunity to focus more on the material. They were also given an thorough explanation of the material by the conservator Old Printed Books of the KB.

- The testers might have lost interest after a number of sessions and thus lose some of the speed they work with.

- Not all tools have the same functionalities. For example, it is possible to adjust the segmentation in Alto Edit and LMU's tool, but not in PlaIR. The choice was made to only change the segmentation in cases where a word was split up, as this would have been counted as an error in the evaluation. Words which were fused were corrected by taping a space, but not by correcting the actual segmentation. This might lead to a slower speed of correcting, but ultimately to a higher accuracy.

- The re-OCRing with ABBYY FRE 10 and the historical dictionary does not require human input. The same goes for re-OCRing with Adaptive OCR.

- It is possible that the setup is disadvantageous to the batch corrections of LMU's tool, which works best on complete books. To counteract this, larger sets were used in the pilot than originally planned.

- The OCR files (ABBYY XML) needed for the LMU tool were not available in the KB. ABBYY SDK 9 is thus used to generate the needed XML files. A comparison between the input and output of the tool can be seen in paragraph 4.3.

- The PlaIR platform is originally designed to work with newspapers, which might be disadvantageous for the users. The University of Rouen made some small adjustments to the platform to make correcting books easier.

- The setup of the pilot contradicted with the way CONCERT works and the correction had to be stopped at a crucial moment. IBM chose to process the books themselves to make a comparison possible. As this meant working for a longer period of time on the material than the other tools, this has been added in a separate paragraph. Please note that the sessions will probably have taken longer as the person correcting the books is unfamiliar with Dutch.

4. Results

As the KB was interested to see how much improvement was possible with regards to the current OCR, these were taken as the benchmark. All texts have been converted into plain text to by-pass the various output formats and to ease the evaluation. For an overview of the various in- and outputs per tool, please see Appendix A. Only word accuracy has been done for the evaluation, as this is most important to the user when searching a collection.

It is important to note that the evaluation set has undergone an automated correction to convert all ligatures and other special characters into plain text. The ligatures have been split into regular characters (e.g. æ → ae) and all various Unicode hyphens have been converted into a standard hyphen (e.g. -). The other sets have not been corrected, as OCR does not yet recognise ligatures or other special characters and the testers had been given the instruction to not use any special characters and to separate all ligatures.

The evaluation was done using the IMPACT NCSR OCR Evaluation workflow[14] and the INL Word Evaluation tool[15] via the IMPACT Framework. All workflows used will be made available on myExperiment.

4.1.1 Set 1-1

Set 1-1 has been corrected by the following testers:

Alto Edit	Person 1
LMU	Person 3
PlaIR	Person 4

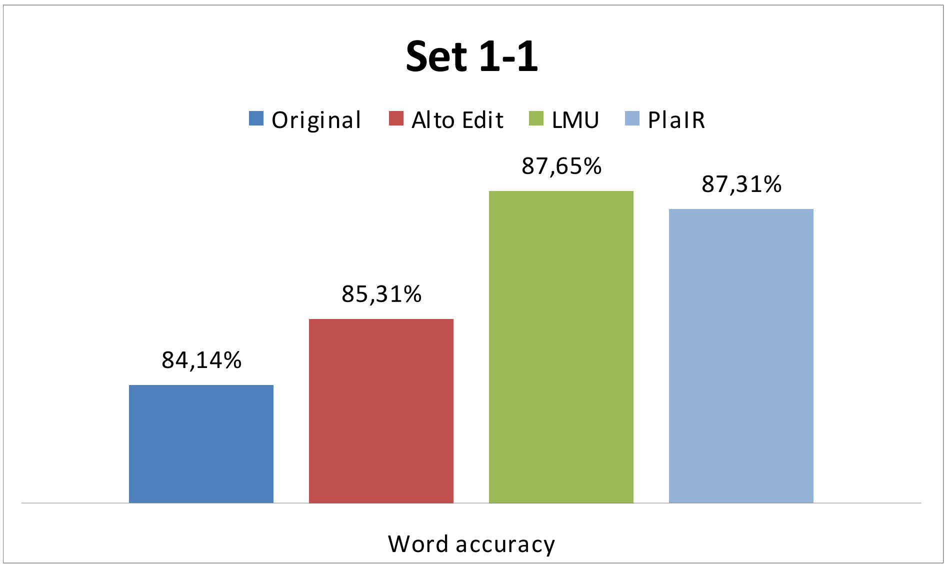


Figure 12 – Evaluation Set 1-1

[14] http://www.myexperiment.org/workflows/1888.html

[15] http://www.myexperiment.org/workflows/2555.html

As can be seen in Figure 12, the original OCR has a word accuracy of 84.14%. After the correction with Alto Edit, there is an increase to 85.28%. The biggest increase in word accuracy is the output of LMU's tool, which goes up to 87.65%. PlaIR comes very close to this result with a word accuracy of 87.31%.

It is surprising to see that LMU and PlaIR are almost at the same level, as PlaIR does not use any automated correction, while LMU's tool does. This difference could be accounted to the fact that PlaIR did use the original ALTO as input, while LMU's tool required newly generated OCR files. For a comparison of the input and output of LMU's tool, please see paragraph 4.3.

4.1.2 Set 1-2

Set 1-2 has been corrected by the following testers:

Alto Edit	Person 2
LMU	Person 1
PlaIR	Person 5

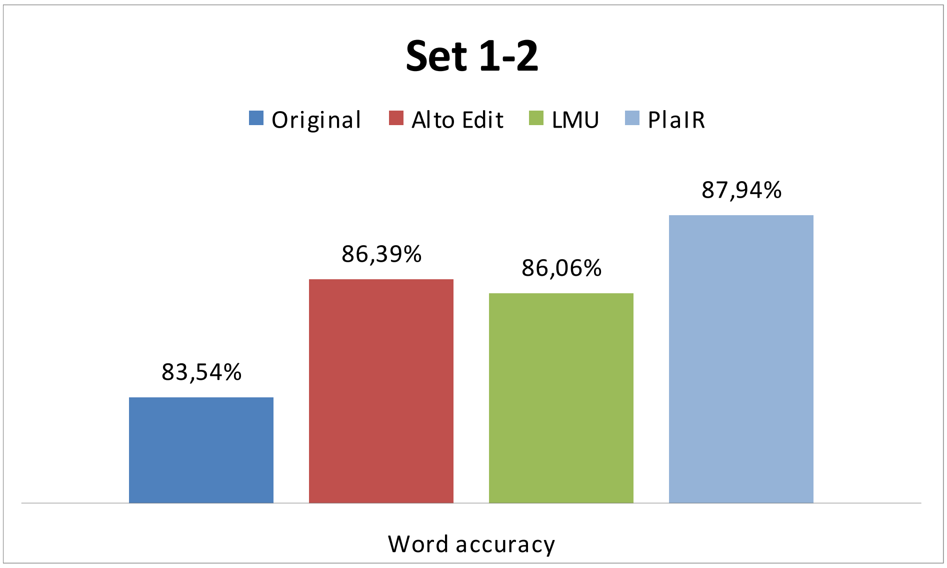


Figure 13 – Evaluation Set 1-2

As can be seen in Figure 13, working with PlaIR provides the best results in this set, followed by Alto Edit and LMU's tool. It is worth knowing that the tester working with the PlaIR tool corrected almost 50% more lines than the other testers (see paragraph 4.5).

It is striking that the two basic tools, Alto Edit and PlaIR, have a reasonably high score in this set compared to the more advanced tool. This difference might again be contributed to the use of the original ALTO files in the two higher scoring tools (see paragraph 4.3).

4.1.3 Set 1-3

Set 1-3 has been corrected by the following testers:

Alto Edit	Person 3
LMU	Person 2
PlaIR	Person 6

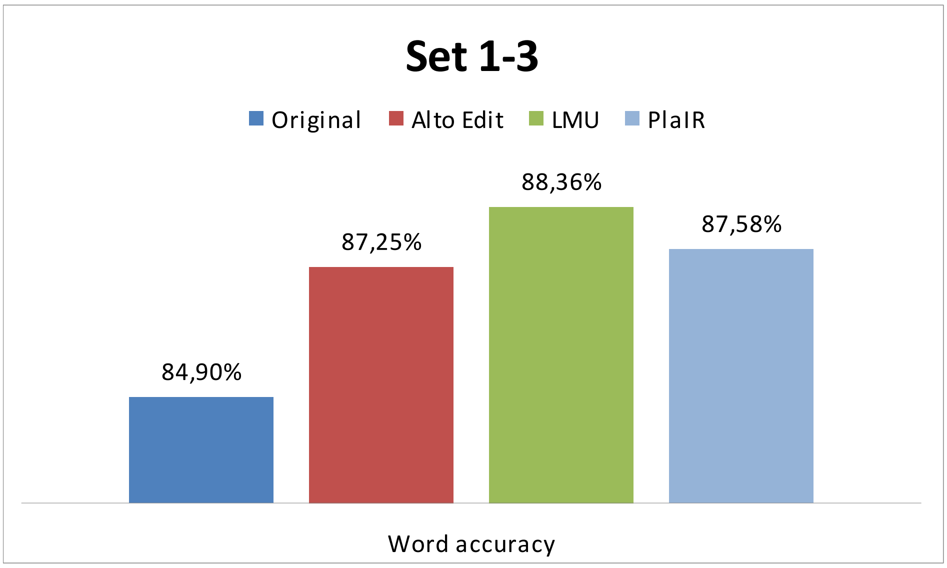


Figure 14 – Evaluation Set 1-3

As can be seen in Figure 14, the top result for Set 1-3 is when using the LMU Profiler and Post Correction tool. PlaIR and Alto Edit have an almost equal result compared to the original ALTO quality.

4.1.4 Book 1

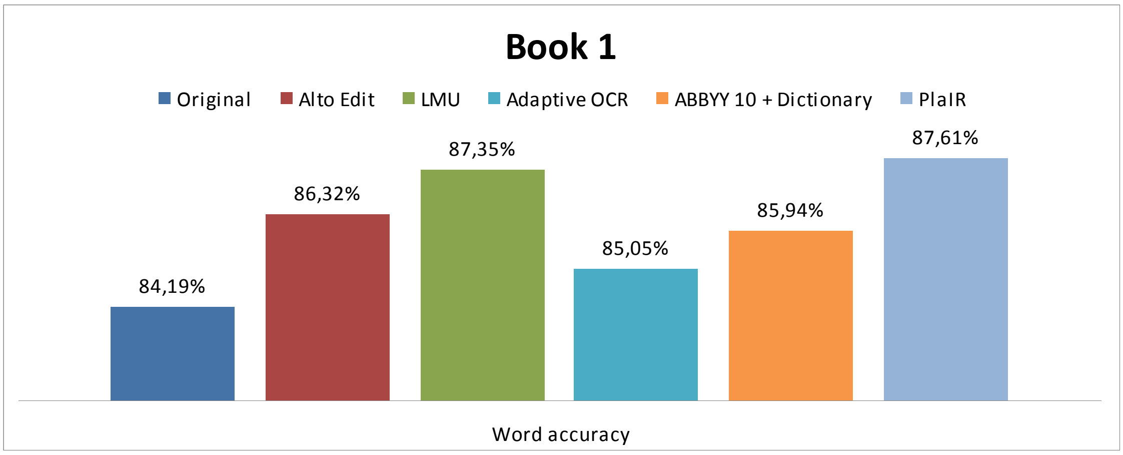


Figure 15 – Evaluation Book 1[16]

For this evaluation, all sets have been combined to calculate the results for the entire book. This evaluation also includes the re-OCRed whole book, ABBYY10 with historical dictionary and IBM's Adaptive OCR. Please note that

[16] This comparison combines tools that include manual correction of the text and tools that processes the material automatically without any human interference. Please see the Conclusion for a more elaborate explanation of the differences between the various methods.

changes to the implementation of the historical dictionary were made during this pilot, which provide better results. Please see the Extension Report by INL for more information.

Compared to the OCR currently available on the EDBO website, when using post-correction tools, correcting with the LMU Profiler and Post Correction Tool and the PlaIR platform provides the best results, followed by Alto Edit. For the re-OCRed book, ABBYY10 with historical dictionary proves to be most successful, followed by IBM's Adaptive OCR. When combining the Adaptive OCR with the CONCERT platform, a higher accuracy can be achieved (see paragraph 4.4).

It is remarkable that LMU's tool and PlaIR are very close together, since the functionalities of the tool are very different. LMU's tool might provide an even better result when starting from the current OCR files instead of the newly generated ones that were required in the pilot (see paragraph 4.3). Starting post-correction from a re-OCRed output would of course be ideal, as this would give firstly a higher word accuracy from the new OCR, which would mean less work in post-correction.

4.2.1 Set 2-1

Set 2-1 has been corrected by the following testers:

Alto Edit	Person 4
LMU	Person 6
PlaIR	Person 1

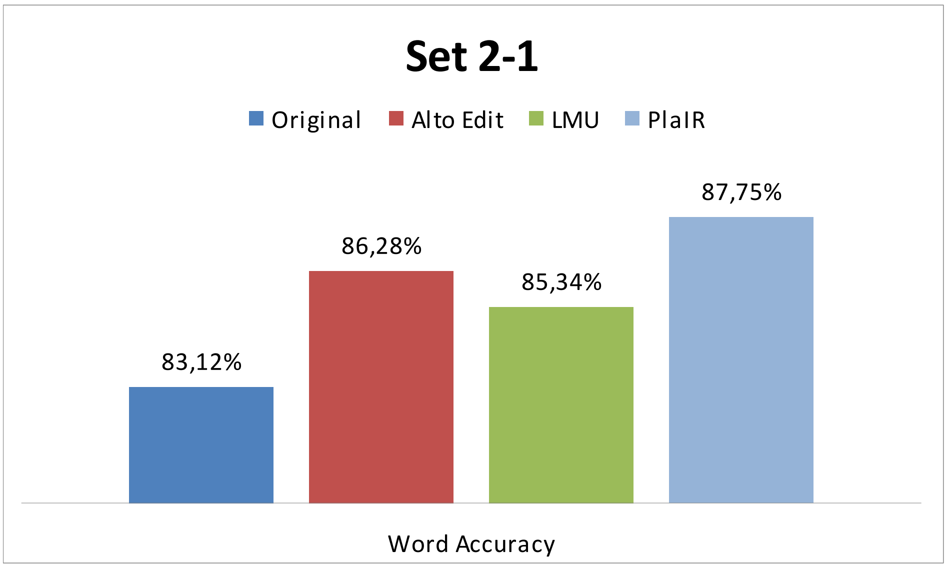


Figure 16 – Evaluation Set 2-1

As can be seen in Figure 16, LMU has scored quite low in this Set, while PlaIR comes out very high with an increase of more than 4,5%. This difference might be attributed to the use of the original ALTO files or possibly even to the testers.

4.2.2 Set 2-2

Set 2-2 has been corrected by the following testers:

Alto Edit	Person 5
LMU	Person 4
PlaIR	Person 2

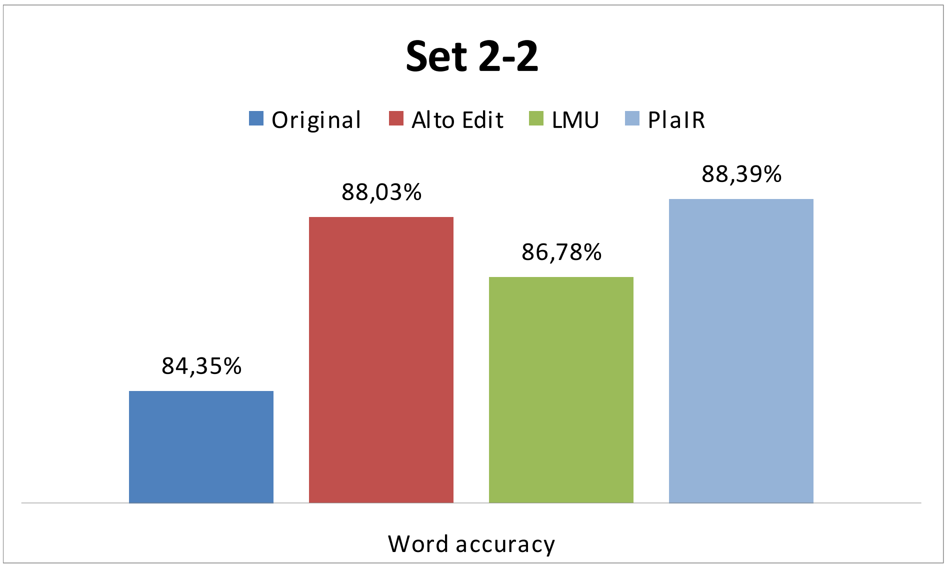


Figure 17 – Evaluation Set 2-2

As with Set 1-1, LMU has the lowest increase and PlaIR the highest. The results of correcting with Alto Edit are almost equal to those of PlaIR. Once again, the result of LMU's tool might be higher when using the original ALTO files (see paragraph 4.3).

4.2.3 Set 2-3

Set 2-3 has been corrected by the following testers:

Alto Edit	Person 6
LMU	Person 5
PlaIR	Not tested

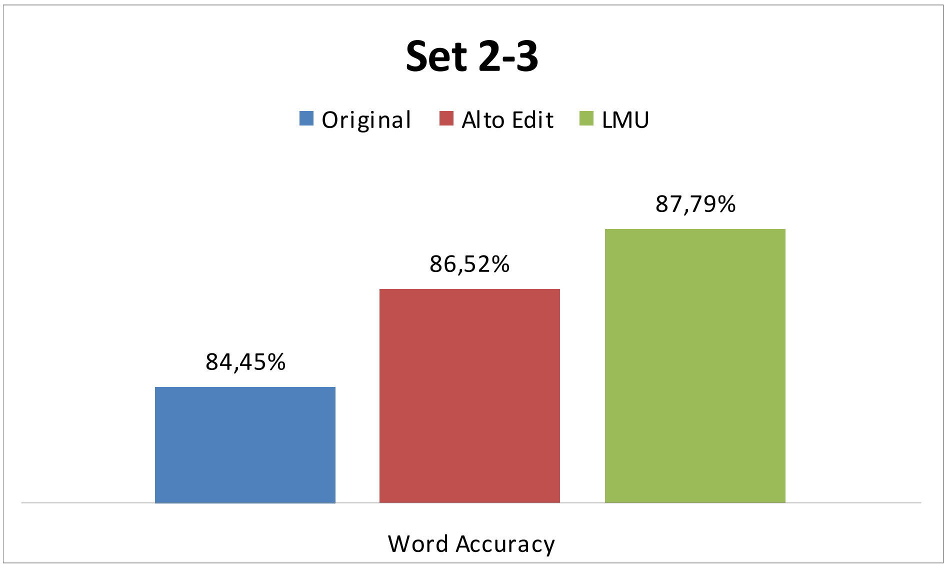


Figure 18 – Evaluation Set 2-3

For the final pilot day, one of the testers was not available. We thus chose to test one set less with the PlaIR platform. As can be seen in the figure above, Set 2-3 has the highest word accuracy after correction with LMU's tool, followed by Alto Edit.

4.2.4 Book 2

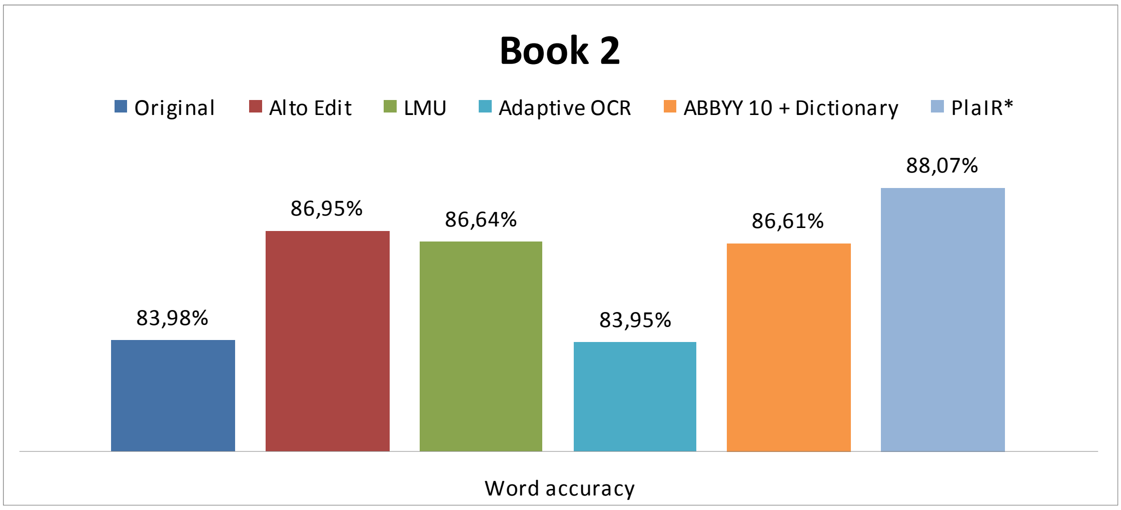


Figure 19 – Evaluation Book 2[17]

For this evaluation, as in paragraph 4.1.4, all sets were combined to make a comparison of the complete book possible. Re-OCRing is thus also included in this evaluation, as there is no need for specific sets when automatically processing a book.

When looking at Book 2 in Figure 19, it can be seen that quite an improvement can be made on the quality of the OCR. When using a post correction tool, the PlaIR platform produces the highest accuracy[18], followed by Alto Edit and the tool by LMU. As LMU did not begin with the original ALTO, but with the lesser quality XML files it is very well possible that an even higher rate can be achieved when taking the original OCR files as input (see paragraph 4.3).

Processing the whole book again with a different OCR method also provides some improvements. Adaptive OCR is almost on the same level as the original OCR, but combining ABBYY10 with a historical dictionary proves to be very successful with a word accuracy of 86.61%. During this pilots, the implementation of the dictionary was updated and produces even better results. For more information, see the INL Pilot Report.

[17] This comparison combines tools that include manual correction of the text and tools that processes the material automatically without any human interference. Please see the Conclusion for a more elaborate explanation of the differences between the various methods.

[18] Please note that this evaluation excludes Set 2-3, as that was not tested in the pilot.

4.3 LMU Post Correction tool

As it was not yet possible to work with the ALTO files from the KB, the LMU Post Correction tool was set up using ABBYY 9 XML files. This difference between data might account for the variations in results.

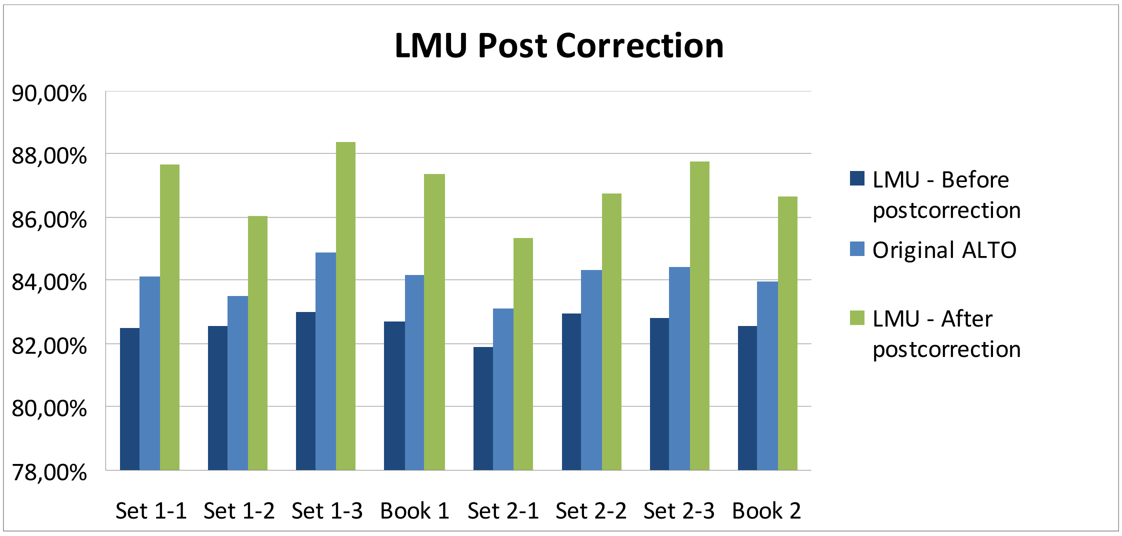


Figure 20 – LMU Post Correction Tool

As can be seen in Figure 16, the quality of the OCR before post correction was lower than the current OCR quality. The increase in accuracy would thus be even greater when using the ALTO files produced in the project as input. However, this is unfortunately not (yet) possible.

4.4 CONCERT

Due to the setup of the pilot, CONCERT was not tested in the way it was intended to be used, resulting in misrepresentative outcomes. The tool is developed to work with input from a large number of volunteers on a large selection of material. The processing time between sessions can take some time, which means a certain set is not available. When working on multiple sets with multiple people, this is not a problem, as a volunteer can simply work on another set.

However, with the setup of the pilot, this was not possible. During the pilot it indeed became a problem that some sets were not available for a longer period of time. The correction of the OCR thus had to be stopped at a crucial stage in the process of CONCERT, making the output incomparable to other correction methods. However, a comparison could be made between the Adaptive OCR software developed by IBM and ABBYY 10 (see paragraph 4.1.4 and 4.2.4). After the pilot, IBM has chosen to process the books completely using CONCERT, thus making a general comparison possible.

The material was processed starting from the outputs of the Adaptive OCR. One person worked on a whole set, which took about four hours to complete. Please note that this tester is not familiar with Dutch, which might slow down the process. However, the person was familiar with the tool and thus needed no training. As there was no time restraint for this test, the outcomes cannot be compared to the other tools.

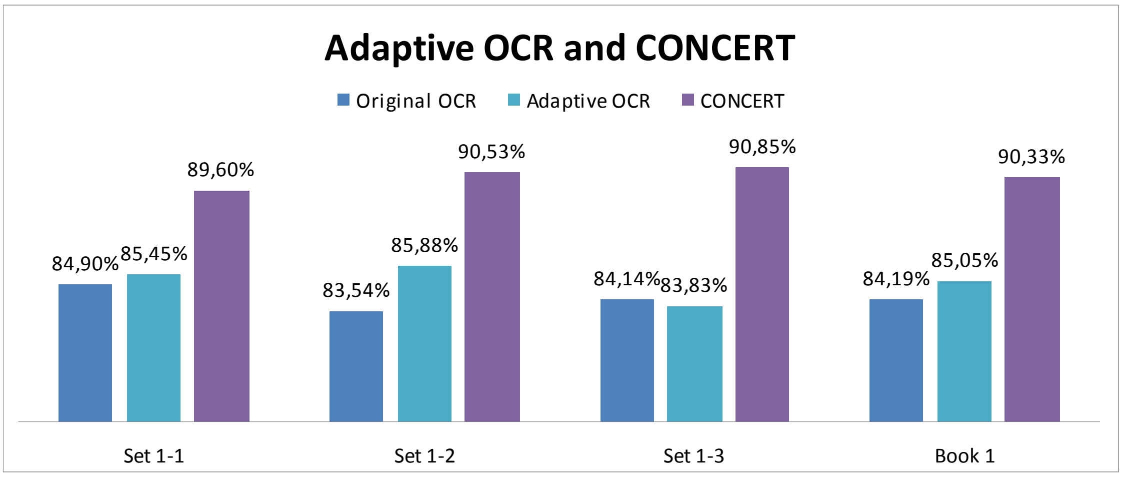


Figure 21 – Adaptive OCR and CONCERT

As can be seen in Figure 21, using CONCERT in combination with the Adaptive OCR provides quite an improvement on the original OCR quality ending up with a 90,33% word accuracy on Book 1. This shows that when this tool is used by a skilled volunteer, the eventual results are very high.

4.5 Speed of corrections

Within the PlaIR platform it is possible to see who has corrected the most lines of text. This gives an indication of the speed the testers worked with. As can be seen in the table below, most testers have corrected around 1000 lines, while Person 5 has corrected almost 1500 lines.

Person 5	Set 1-2	1450 lines
Person 4	Set 1-1	1094 lines
Person 2	Set 2-2	1089 lines
Person 1	Set 2-1	1078 lines
Person 6	Set 1-3	1041 lines

The LMU Post Correction system also has the possibility of saving the data with a timestamp. This information was collected and put into a graph. The Users in the graph are numbered differently than in the pilot:

Person 1	User6	Set 1-2
Person 2	User3	Set 1-3
Person 3	User2	Set 1-1
Person 4	User1	Set 2-2
Person 5	User5	Set 2-3
Person 6	User4	Set 2-1

As can be seen in Figure 21, User 2 (is Person 3) has made the most corrections and User 1 (is Person 4) the least with a difference of 832 corrections overall.

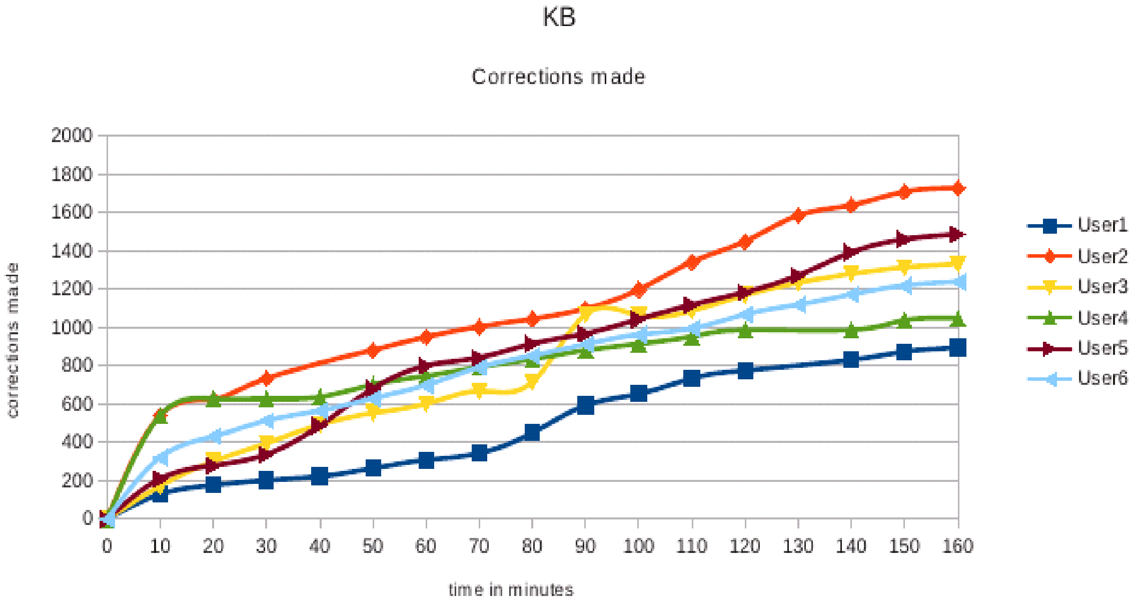


Figure 22 – Corrections with the LMU Post Correction Tool

5. Conclusion

As mentioned before, it is very difficult to do a cross-comparison between tools which vary significantly in functionality and setup. Due to this, it is important that a library or institution that wants to improve OCR has a clear purpose of the project and the goals they wish to achieve.[19]

Looking at the choices to be made, the tools can be coarsely divided into three categories:

Basic tools	Advanced tools	Re-OCRing
Alto Edit	LMU Profiler and Post correction tool	ABBYY FRE 10 with a historical Dutch dictonary
PlaIR	CONCERT	Adaptive OCR

5.1 Basic tools

When a library or institution wants to involve the general audience in the OCR correction, the basic tools are to be recommended. These are easy to use, can be accessed via a website, but require some effort of the library for maintenance and quality control.

Of the two basic tools used in this pilot, the PlaIR platform[20] provides the best results compared to the original OCR. This is not surprisingly, as Alto Edit is a tool developed with little effort within a small research project, in comparison to the PlaIR platform, which stems from the Trove tool developed in Australia, and has been adapted by the University of Rouen. However, both tools do provide quite an improvement on the original OCR.

Both tools are available open source.

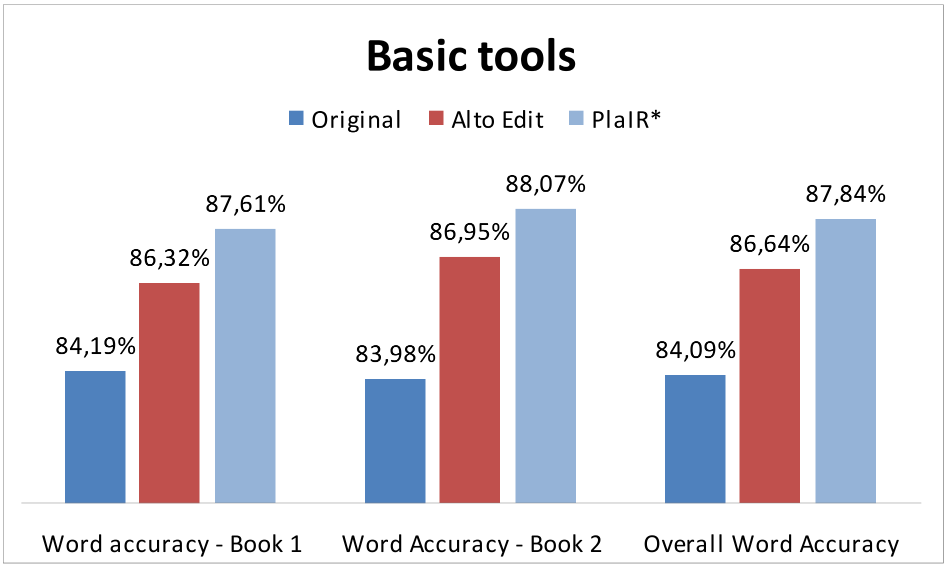


Figure 23 – Basic tools

[19] This conclusion compares the output of the tools of the current OCR files produced for the KB. When an institution does not have OCR files for a collection and wishes to use either one of the tested tools, it would be wise to do a separate pilot, as the results can be quite different for new collections.

[20] Please note that for Book 2, Set 2-3 is not included in this evaluation, as it was not tested in the pilot.

5.2 Advanced tools

The advanced tools both have a different setup, but offer much more functionalities than the basic tools. Because of these elaborate options, the tools require effort from more experienced users and possibly even one or more staff members. Both tools in this category unfortunately did not have the possibility to work with the original ALTO files, which may give a distorted result when comparing to the current quality of the OCR.

From the two advanced tools tested in the pilot, only LMU's Profiler and Post Correction Tool was evaluated as part of the pilot. It is quite possible that the results of LMU's tool will be even higher when using the original ALTO files as input (see paragraph 4.3). The CONCERT tool was processed separately without a time limit and can thus not be compared to the outcomes of the other tools (see paragraph 4.4 for results).

Both tools require a (partial) license.

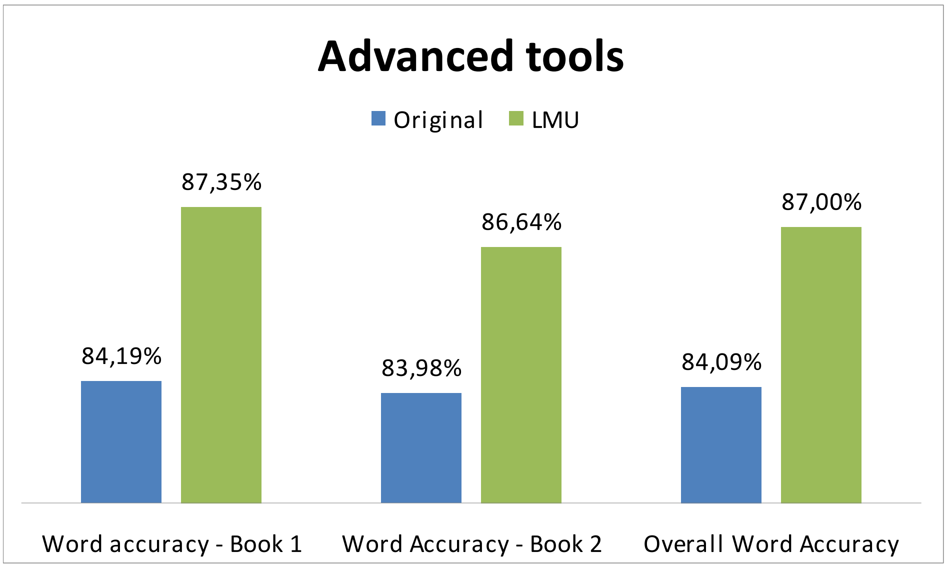


Figure 24 – Advanced tools

5.3 Re-OCRing

When a library or institution wants to use as little effort as possible to improve the quality of their OCR, re-OCRing is the best option. This only requires the master images and the rest of the process is automated. For Book 1, this already provides a 1,75% improvement on the current OCR quality when using ABBYY 10 and a historical Dutch dictionary and a 2,63% increase for Book 2. The results shown here for ABBYY10 were done with a very basic setup of the tool and might be significantly improved when adapting the system to the material.

Use of ABBYY FRE 10 and the historical dictionary requires a license for both parts. The Dutch historical dictionary has been developed by the Instituut voor Nederlandse Lexicologie.

Another possibility of re-OCRing is with IBM's Adaptive OCR. These outputs can also be used with the web-based postcorrection system CONCERT, providing volunteers with the possibility of correcting the OCR and thus combining re-OCRed material with post-correction. For Book 1 this would mean a start of 85.05% word accuracy instead of the currently available 84.19%. For the results when combining the Adaptive OCR with CONCERT, see paragraph 4.4.

Use of the Adaptive OCR and CONCERT requires a license.

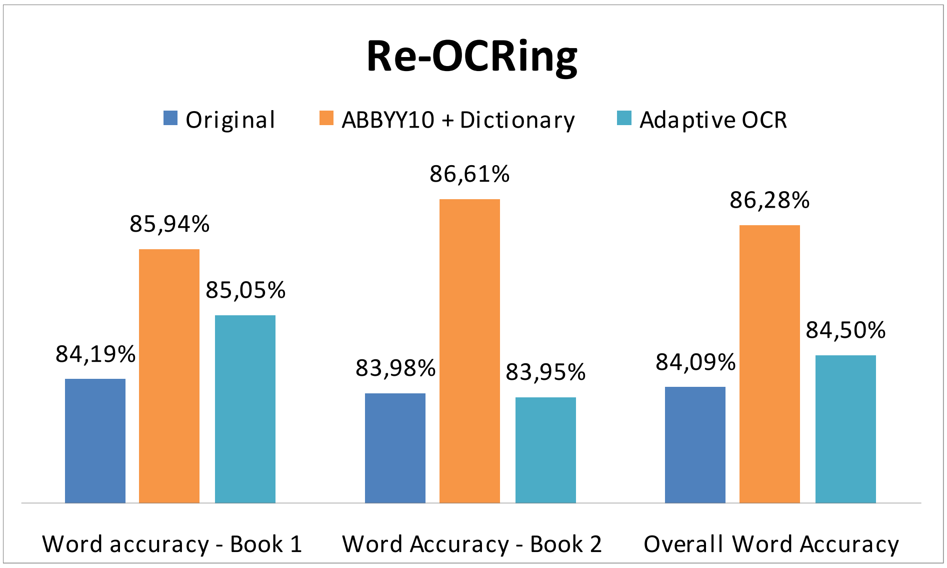


Figure 25 – Re-OCRing with a historical dictionary

5.4 Conclusion

To conclude, the goal of any OCR improvement project is very important when choosing a specific tool or method. Depending on the resources available and wishes of an institution, any of the tested tools may provide the desired output and improved OCR results. Ideally, all material is re-OCRed before any post correction, resulting in the highest possible word accuracy.

This report only gives an insight into a very specific use case. It is to be recommended that any library or institution do their own evaluation before choosing a specific solution, as this will result in the best possible combination for the type of material and goals set.

6. Recommendations

- It is very important for an institution to have a very clear goal in mind when setting up an OCR correction project. For instance, when the library wants a very approachable crowd sourcing setup, it is better to choose one of the basic tools. However, when more effort is possible from within the library to train and support any users, it is recommendable to use one of the more advanced tools, as they could result in a much better quality. When the library wants to use as least effort as possible, running the OCR process again with a historical dictionary is the best option, as it provides good results with very little effort.

- The tools that can use the current ALTO files as an input give very good results. It is possible that the other tools, where this is not (yet) possible, produce even better results when starting from the original OCR. It would be very interesting to research this further. If this is not possible, using a tool which can work with the library's input may provide better results.

- It is important to investigate what the tool requires as input and provides as output and how this fits into the library's internal workflow. If, for instance, the library uses ALTO files and the tool can only provide plain text files as an output, some functionalities, such as highlighting search terms will not be possible anymore. An overview of all tested tools can be found in Appendix A.

- Some tools are only available with a specific license. When investigating OCR post correction further, it is recommendable to also take this into account. Since some of the developers were still working on their business model when this pilot was planned, this was left out of scope.

Appendix A : Tool Overview

Tool name	Tool type	Input	Output	Preprocessing	Postprocessing	Options
AltoEdit	• Web-based • Open source • Developed by KB	• ALTO • Master images (JP2)	ALTO	Rule based correction of long s	None	• Correct segmentation • Search-and-replace
LMU Profiler and Post Correction Tool	• Stand-alone • Linux • Profiler is a service via LMU • Post-correction GUI is open source	• ABBYY XML • Tiff images	Plain text	None	Plain text files need to be separated per page.	• Correct segmentation • Batch corrections according to specific profile
CONCERT	• Web-based • Can run at institution or via IBM • License via IBM	• Tiff images • Optional: OCR with coordinates on a character level	ALTO	When no character level OCR is available, the tool uses IBM's Adaptive OCR to produce it.	None	• Correct segmentation • Character level corrections • Internal dictionary
PlaIR	• Web-based • Based on TROVE • Open source	• OCR files • Tiff images	Plain text (will be updated to ALTO)	None	None	None
ABBYY FRE10[21] + Dictionary	• Automated process • License via Instituut voor Nederlandse Lexicologie and ABBYY	Tiff images	PAGE XML	None	None	None
Adaptive OCR	• Automated process • License via IBM	Tiff images	ALTO-Ex	None	None	None

[21] For more information about the in- and outputs of ABBYY, please see http://www.abbyy.com/ocr_sdk_windows/technical_specifications/io_formats/