# Europeana Newspapers OCR Workflow Evaluation[†]

Stefan Pletschacher, Christian Clausner and Apostolos Antonacopoulos

Pattern Recognition and Image Analysis (PRImA) Research Lab
School of Computing, Science and Engineering, University of Salford, Greater Manchester, United Kingdom
http://www.primaresearch.org

## ABSTRACT

This paper summarises the final performance evaluation results of the OCR workflow which was employed for large-scale production in the Europeana Newspapers project. It gives a detailed overview of how the involved software performed on a representative dataset of historical newspaper pages (for which ground truth was created) with regard to general text accuracy as well as layout-related factors which have an impact on how the material can be used in specific use scenarios. Specific types of errors are examined and evaluated in order to identify possible improvements related to the employed document image analysis and recognition methods. Moreover, alternatives to the standard production workflow are assessed to determine future directions and give advice on best practice related to OCR projects.

## Categories and Subject Descriptors

H.3.7 [INFORMATION STORAGE AND RETRIEVAL]: Digital Libraries, I.7.1 [DOCUMENT AND TEXT PROCESSING]: Document and Text Editing --- Document management, I.7.5 [DOCUMENT AND TEXT PROCESSING]: Document Capture, I.5.4 [PATTERN RECOGNITION] Applications --- Text processing, Computer vision.

## Keywords

OCR, performance evaluation, newspapers, historical documents

## 1. INTRODUCTION

Comparatively little historical content has been digitised so far [1]. To address this, several digitisation projects are under way in major libraries and archives around the world. However, in the great majority of cases, the results of digitisation projects are scanned pages with no encoded textual content. In the remainder of cases where there is “full text” available, this is by far simply the result of applying an OCR system to the scanned pages, without any further correction of either the text or the layout structure.

It is of crucial importance to objectively evaluate the results of digitisation projects not only in terms of apparent accuracy (e.g. percentage of correct words) but more importantly in the context of their different intended use scenarios also. Knowing the results of such an evaluation, scholars can manage their expectations of the material and content holding institutions can better understand and improve the quality of their collections. Of equal importance is also the evaluation of the OCR workflows that produced those results in order to help researchers to improve them, and content holders to manage their expectations and, consequently, their digitisation priorities.

The need for quality evaluation of large-scale digitisation project results has been recognised for some time. Early attempts used *character recognition rate* and *OCR confidence rate* [2] as measures. The merits of using more meaningful measures such as *word recognition rate* were, subsequently, argued [3] (scholars use words as search terms, not characters). However, those evaluation approaches still do not go far enough to provide a meaningful measure of quality in real-world use scenarios (e.g. phrase search) and are very laborious as they involve significant amounts of manual work. A new comprehensive evaluation framework was created by the authors and used extensively during the IMPACT project [4] and adapted for historical newspapers for the Europeana Newspapers project [5] (see below). It comprises semi-automated tools for ground truthing and scenario-based performance analysis. The framework has also been used successfully in a number of international competitions (e.g. [6][7][8]).

This paper describes the process and analyses the findings of the evaluation of the OCR workflow (and corresponding results) of the Europeana Newspapers project [5], one of the largest-scale OCR/refinement projects and the most significant contributor of content to the European Library [9] so far. In addition, alternatives to the main workflow are examined and evaluated to assess possible improvements and determine future directions.

To the best of the authors’ knowledge this is the first time that such a comprehensive analysis of the performance of a large-scale OCR production workflow is performed and presented for the benefit of document analysis researchers, content holding institutions and humanities scholars.

In the following section, the background and objectives of the Europeana Newspapers project are summarised to provide the context for the rest of the paper. Section 3 describes the evaluation infrastructure, detailing also the different evaluation metrics and scenarios. The evaluation results are presented and analysed in Section 4, where alternative workflows are also discussed and evaluated. Finally, Section 5 concludes the paper with remarks on the overall experience and results.

## 2. THE EUROPEANA NEWSPAPERS PROJECT

The Europeana Newspapers project [5] was an EU-funded Best Practice Network under the theme CIP-ICT-PSP.2011.2.1 - Aggregating content in Europeana. The project lasted from 1 February 2012 until 31 March 2015 and brought together 18 project

*HIP '15*, August 22, 2015, Nancy, France
© 2015 ACM. ISBN 978-1-4503-3602-4/15/08…$15.00
DOI: http://dx.doi.org/10.1145/2809544.2809554

[†] This work was funded through the EU Competitiveness and Innovation Framework Programme grant Europeana Newspapers (Ref. 297380).

**Table 1. Evaluation profiles for use scenarios**

| Keyword search in full text | Phrase search in full text | Access via content structure | Print/eBook on demand | Content based image retrieval |
|---|---|---|---|---|
| • Only text regions are of interest<br>• Miss of regions or parts of regions is penalised most<br>• Splits are penalised only a little (a keyword may have been split)<br>• Merges are less important (only merges across columns may be problematic when hyphenation is involved)<br>• Misclassification from text to text is irrelevant (e.g. paragraph misclassified as heading)<br>• False detection is irrelevant (additional regions are unlikely to compromise the indexing)<br>• Reading order is ignored (only the occurrence of words is of interest, no matter in which order)<br>• Bag of Words evaluation for text is sufficient | • Only text regions are of interest<br>• Miss of regions or parts of regions is penalised most<br>• Merge of regions not in reading order is highly penalised<br>• Merge or split of consecutive text blocks ('allowable') only minimally penalised<br>• Splits are unwanted but not especially emphasized (default penalty)<br>• False detections are disregarded<br>• Focus on word accuracy for text evaluation (high accuracy required) | • Access via content structure<br>• Main focus on textual elements<br>• Special emphasis on subtypes headings, page numbers and TOC-entries<br>• Miss, partial miss and misclassification is penalised most<br>• Merge and split of regions not in reading order is highly penalised<br>• Merge and split of consecutive text blocks ('allowable') only minimally penalised<br>• False detection penalised least<br>• Reading order important<br>• Focus on word accuracy (moderate to high requirements on text accuracy) | • Text regions are considered more important than other regions<br>• Miss of regions or parts of regions is penalised most<br>• Merges get a high penalty<br>• Merge or split of consecutive text blocks ('allowable') only minimally penalised<br>• Image, graphic and line drawing are treated as equal (misclassification not penalised)<br>• Noise and unknown regions are irrelevant<br>• Reading order important<br>• Focus on word accuracy (moderate to high requirements on text accuracy) | • Only image, graphic, line drawing and chart are of interest<br>• Image, graphics, line drawing and chart are considered one class (no misclassification error)<br>• Miss, partial miss and misclassification are penalised most<br>• Reading order and allowable splits, merges are disregarded<br>• Low requirements on text accuracy (only captions) |

partners (mostly national/major libraries in Europe), 11 associated partners and 35 networking partners to achieve the ambitious goal of making vast amounts of European digital historic newspapers available via two prominent cultural heritage websites, Europeana [10] (search through metadata) and The European Library [9] (full-text content searchable). Over 11 million newspaper pages were processed through the OCR workflow and an additional 2 million pages were processed through a separate semi-automated layout analysis workflow, which also identified articles etc. The full content produced by the project is freely searchable and accessible to the general public.

The OCR workflow was run at the University of Innsbruck on a number of servers working around the clock for several months. The workflow steps were decided upon based on experience and experimental validation.

The scanned newspaper pages were first catalogued (recording language and text metadata among other things), organised into appropriate folder structures (corresponding to individual newspaper titles and issues) and binarised (to significantly reduce the amount of data to be transferred) at each of the content-holding institution sites. Due to the very significant volume of data to be shipped, it was faster to send those pre-processed images on hard disk drives by mail than transfer through the Internet. Once at the University of Innsbruck, the images and related information were validated (using a number of tools developed by the project) into individual newspaper titles and issues, and sent to the OCR process with parameters appropriate to each image according to the metadata recorded. ABBYY FineReader Engine 11 SDK [11] was used as the OCR engine due to its superior performance and flexibility in configuration. The results of OCR for each page were exported to ALTO format [12] and organised with a METS [13] structure.

# 3. EVALUATION INFRASTRUCTURE

Efficient and reproducible evaluation of large-scale OCR projects requires a number of resources as well as tools for automation to be put in place. In the following, the general evaluation infrastructure as it was used in the Europeana Newspapers project (and which can in principle also be applied to other projects) is presented.

## 3.1 Use Scenarios

The motivation of scenario-based evaluation comes from the observation that abstract error metrics need to be put in context of the intended use in order to obtain meaningful scores. Very typical examples which highlight this are *keyword search* and *phrase search* in full text. While both rely on text recognition results to be of sufficient quality, phrase search has far greater requirements in terms of the layout needing to be recognised correctly as well. For instance, if two columns on a newspaper page were erroneously merged, the individual words would still be accessible for keyword search but phrase search would fail on any portions of the text now wrongly spanning the two merged columns rather than following the line breaks within each individual column.

In order to identify use cases that were relevant to the partner libraries and the material in Europeana Newspapers, a survey was carried out resulting in five use scenarios which were to be considered in the final evaluation. Accordingly, the second part of the evaluation section is based on the evaluation profiles described in Table 1, representing settings and error weights corresponding to the five use scenarios.

## 3.2 Metrics

Each scenario defines an evaluation strategy that includes settings and weights which are applied to the specific metrics resulting from the comparison of OCR output and ground truth. As such, metrics can be seen as qualitative and/or quantitative measures for certain types of errors exhibited by the OCR result. In the following the main metrics which were used for performance evaluation are described.

### *3.2.1 Text-based evaluation*

The idea behind all text-based evaluation methodologies is to compare the OCR result text (e.g. Abbyy FineReader output) against the ideal text (ground truth). Depending on the level of detail required by the use scenario different text comparison approaches can be used.

A basic metric is *word accuracy* which requires a serialisation of the result and ground truth text and then measures word by word how well the two strings match [14]. It calculates how many edit, delete, and insert operations are required to make one text equal to another. It is important to note that this metric is sensitive to the order of words.

The same principle can also be applied to *character accuracy* only that, instead of edits, deletes, and inserts of whole words, the character level is used. However, due to the nature of the algorithm [14], calculating the character accuracy is too resource intensive for long texts (such as found on newspaper pages). Moreover, character accuracy is typically only interesting to developers of OCR systems and not normally used to assess the suitability of recognised documents for specific use scenarios.

The need to handle text serialisations of potentially very long documents (which is more often than not the case for newspapers) leads to the so called *Bag of Words* metrics which do not take into account the order of the words in the texts. Only the fact whether an OCR system recognised words correctly or not is of significance. There are two flavours of this measure: For the *index based* success rate it is only important for the OCR engine to find each word at least once and not to introduce false words. The *count based* success measure is stricter and demands the correct count of recognised words (e.g. have all occurrences of the name Shakespeare been found or only seven out of nine).

Although similar, both success rates may differ significantly on the same document due to the specific focus of each.

### *3.2.2 Layout-based evaluation*

In addition to textual results, page reading systems are also expected to recognise layout and structure of a scanned document page. This comprises *segmentation* (location and shape of distinct regions on the page), *classification* (type of the regions defined by the segmentation; e.g. text, table, image, etc.), and *reading order* (sequence/grouping of text regions in which they are intended to be read).

Evaluation profiles specify which of those measures to use and how much impact they should have on the overall result. This includes weights for segmentation errors (merge, split, miss, and false detection), misclassification errors, and reading order errors. Depending on the profile, the overall success rate for an OCR result can vary significantly.

#### *3.2.2.1 Evaluation of segmentation and classification results*

The performance analysis method used [15] can be divided into three parts. First, all regions (polygonal representations of both ground truth and method results for a given image) are transformed into an interval representation, which allows efficient comparison and calculation of overlapping/missing parts. Second, correspondences between ground truth and segmentation result regions are determined. Finally, errors are identified, quantified and qualified in the context of one or more use scenarios.

The region correspondence determination step identifies geometric overlaps between ground truth and segmentation result regions. In terms of Page Segmentation, the following situations can be determined:

- *Merger*: A segmentation result region overlaps more than one ground truth region.
- *Split*: A ground truth region is overlapped by more than one segmentation result region.
- *Miss (or partial miss)*: A ground truth region is not (or not completely) overlapped by a segmentation result region.
- *False detection*: A segmentation result region does not overlap any ground truth region.

In terms of Region Classification, considering also the *type* of a region, an additional situation can be determined:

- *Misclassification*: A ground truth region is overlapped by a result region of another type.

Based on the above, the segmentation and classification errors are *quantified*. This step can also be described as the collection of raw evaluation data. The amount (based on overlap area) of each single error is recorded.

This raw data (errors) are then *qualified* by their significance. There are two levels of error significance. The first is the implicit *context-dependent* significance. It represents the logical and geometric relation between regions. Examples are *allowable* and *non-allowable* mergers. A merger of two vertically adjacent paragraphs in a given column of text can be regarded as allowable, as the result will not violate the reading order. Conversely, a merger between two paragraphs across two different columns of text is regarded as non-allowable, because the reading order will be violated. To determine the allowable/non-allowable situations accurately, the reading order, the relative position of regions, and the reading direction and orientation are taken into account.

The second level of error significance reflects the additional importance of particular errors according to the use scenario for which the evaluation is intended. For instance, to build the table of contents for a print-on-demand facsimile edition of a book, the correct segmentation and classification of page numbers and headings is very important (e.g. a merger between those regions and other text should be penalised more heavily).

Appropriately, the errors are also weighted by the size of the area affected (excluding background pixels). In this way, a missed region corresponding to a few characters will have less influence on the overall result than a miss of a whole paragraph, for instance.

For comparative evaluation, the weighted errors are combined to calculate overall error and success rates. A non-linear function is used in this calculation in order to better highlight contrast between methods and to allow an open scale (due to the nature of the errors and weighting).

#### *3.2.2.2 Evaluation of reading order*

Reading Order describes the sequence in which textual elements on a page should be addressed. It is therefore a key requirement with regard to a document’s logical structure. This information is crucial, for instance, for conversion tasks that need to preserve the original text flow (e-books, PDF, HTML).

OCR results depend strongly on correctly detected reading order, making its evaluation a critical aspect of the overall performance analysis [16]. Ground truth and detected reading order can typically not be compared directly due to differences in region segmentation. Further, complex layouts require a reading order format that goes beyond a simple sequence.

In order to accommodate the requirements specific to newspapers a flexible tree structure with groups of ordered and unordered elements is used. Text elements that are not intended to be read in a particular sequence (e.g. adverts within a page) can have an unordered relation. Objects which may be irrelevant in terms of the actual content (page number, footer etc.) can be left out entirely.

The method employed in the following reduces the influence of differences in segmentation by calculating region correspondences. Partial relations between regions are determined by exploring the reading order trees and are then weighted with the relative overlap of the involved regions. All partial relations for each pair of regions are then penalised according to a certain matrix and are finally combined to a composite penalty [16].

## 3.3 Dataset

The fact that performance evaluation depends on ground truth (representing the ideal result) entails the need for a representative dataset for which these additional resources are available. To this end, a comprehensive and realistic dataset was created during the course of the Europeana Newspapers project. [17].

The dataset was created in three main stages:

- Broad selection and aggregation of representative images and metadata,
- Selection of subsets to be used for evaluation, and
- Production of ground truth for all subsets.

The selection of subsets to be used for evaluating the main production workflow was driven by two major constraints:

1. To narrow the initial selections further down so as to be in line with the available resources (budget).
2. To maintain the representativeness of the individual datasets as far as possible.

It was agreed to fix the size of each subset to 50 images, allowing for reasonable variety while keeping costs within the limits of the budget.

With regard to representativeness it was tried to keep the distribution of languages, scripts, title pages, middle pages, and characteristic layouts as close to the original selection as possible. For practical reasons and to be able to run realistic evaluation scenarios it was also ensured that at least one full issue was included per subset. In total the dataset used for evaluating the main production workflow comprised 600 newspaper pages.

## 3.4 Evaluation Workflow

In order for the evaluation results to be objective and reproducible as well as the overall process to run as efficiently as possible, an automated evaluation workflow was set up, using numerous tools [18] specifically developed for this purpose. Figure 1 shows the overall evaluation workflow.

### 3.4.1 *Ground truth production*

All ground truth data was pre-produced using FineReader Engine 10. Service providers then manually corrected recognition errors (page layout and text). Quality control (assisted by the PAGE Validator tool) ensured ground truth of a predefined accuracy.

### 3.4.2 *OCR result production*

OCR output was produced using the Europeana Newspapers production workflow which included the NCSR Demokritos image binarisation method and Abbyy FineReader Engine 11. The recognition results were obtained in both ALTO XML and FineReader XML format, which were subsequently converted to PAGE XML format [19] to be used by the evaluation tools.

In addition, all document images were also processed with Tesseract, the state-of-the-art open source OCR software, in order to allow comparison of two different OCR engines.

### 3.4.3 *Text-based performance evaluation*

The text recognition performance of OCR systems can essentially be measured by comparing plain text files. For a fair evaluation, the following processing steps needed to be performed:

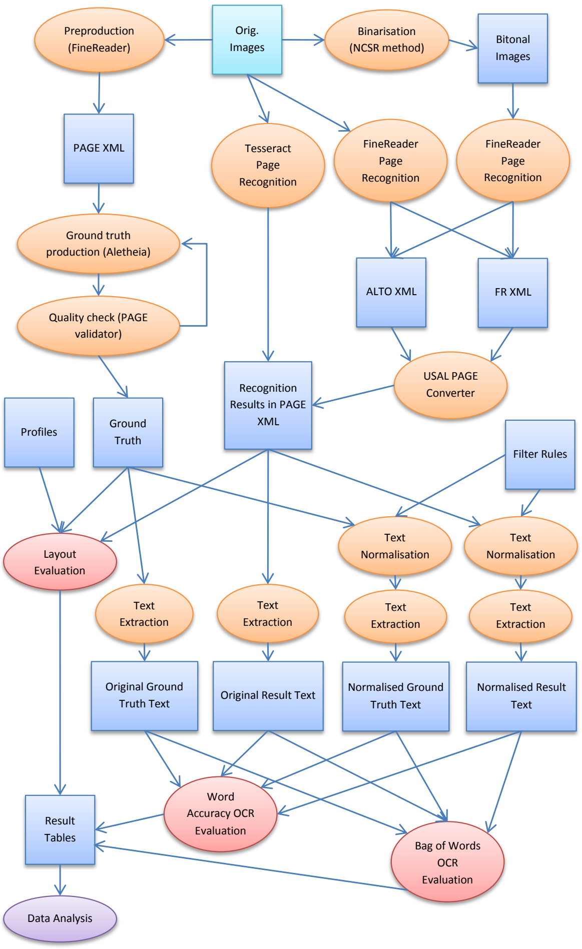
Preproduction (FineReader)
Orig. Images
Binarisation (NCSR method)
Bitonal Images
PAGE XML
Tesseract Page Recognition
FineReader Page Recognition
FineReader Page Recognition
Ground truth production (Aletheia)
ALTO XML
FR XML
Quality check (PAGE validator)
USAL PAGE Converter
Recognition Results in PAGE XML
Profiles
Ground Truth
Filter Rules
Layout Evaluation
Text Normalisation
Text Normalisation
Text Extraction
Text Extraction
Text Extraction
Text Extraction
Original Ground Truth Text
Original Result Text
Normalised Ground Truth Text
Normalised Result Text
Word Accuracy OCR Evaluation
Result Tables
Bag of Words OCR Evaluation
Data Analysis

**Figure 1. Evaluation Workflow**

#### 3.4.3.1 *Text normalisation*

To preserve information, the ground truth text was transcribed as close as possible to the original document. This involved special characters such as the long s or ligatures like ck whenever necessary. For a more realistic evaluation (current OCR engines are still limited with regard to the character sets they can recognise - especially related to historical documents) both ground truth and result text were normalised using replacement rules satisfying the following conditions:

- Characters from private Unicode areas and MUFI (Medieval Unicode Font Initiative) recommendations are always converted or deleted;
- Extremely similar looking characters are mapped to one (all double quotation marks to the standard quotation mark; all single quotation marks to the apostrophe; etc.);
- Ligatures are expanded into individual characters;
- Language specific characters, that look similar in another language, are not replaced (e.g. B in Latin, B [Beta] in Greek, and B [Ve] in Cyrillic).

#### 3.4.3.2 *Text export*

Since the actual Unicode text is embedded in the element hierarchy of PAGE XML files it was necessary to serialise all text streams. To this end, an exporter tool was used for extracting only the textual content into plain text files. This was done for both original and normalised ground truth/result files. This process had to take into account the reading order of text regions so as to arrive at a valid serialisation of the text contained in potentially very complex layouts.

#### *3.4.3.3 Evaluation*

The actual performance evaluation was carried out using the text evaluation tool [18] in two different modes:

- Bag of Words method
- Word accuracy method

For comparison, a total of eight different combinations of input files were processed:

- OCR results based on bitonal and original images
- ALTO XML and FineReader XML format
- Original text and normalised text

### *3.4.4 Layout-based performance evaluation*

All evaluation runs were carried out using the PRImA Layout Evaluation tool [18]. Several factors were taken into account, leading to a total of 20 result tables:

- Five different evaluation profiles matching the use scenarios defined above
- OCR results based on bitonal and original images
- ALTO XML and FineReader XML format

# 4. EXPERIMENTAL RESULTS AND ANALYSIS

This section summarises all the results that were obtained from the evaluation experiments as outlined before. The first part focuses on the performance of pure text recognition (disregarding more sophisticated features like document layout and structure), followed by results based on scenario-driven evaluation (taking into account segmentation, classification, and reading order) in the second part, and aspects related to the choice and configuration of components in the production workflow in part three.

## 4.1 Text-Based Evaluation

Following common practice, the first step towards assessing the accuracy of OCR results is an in-depth analysis on plain text level.

### *4.1.1 Strict*

As indicated before, standard word accuracy is a measure for how well the words contained in two strings match. Since it depends on the respective word order it can be considered a very strict measure.

#### *4.1.1.1 Overall results*

The chart in Figure 2 shows the overall word accuracy for original and normalised text obtained from bitonal images as input and ALTO as output format.

The first observation is that there is a considerable difference between the results based on the original text and on the normalised text. The explanation lies in the fact that current OCR engines are trained only for limited character sets and are typically not designed to recognise special characters which are only common in historical documents (such as the long s) or might be typographical idiosyncrasies. Moreover, one might argue, that recognising the historical variant of a long s as a modern s is what OCR should do in order to allow for meaningful search results on the OCR output. Others, however, would argue that OCR should always return the correct character code, corresponding to the glyph on the page and leave any further interpretation to subsequent systems (such as fuzzy search in information retrieval systems).

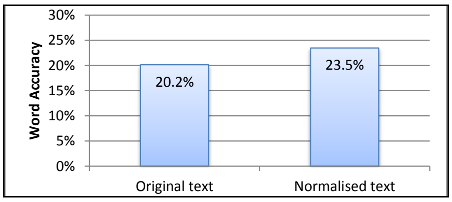

**Figure 2. Overall word accuracy – original vs. normalised text**

The second observation is that the overall accuracy is rather low, even when looking at the relaxed measure based on normalised text. It has to be noted, however, that this is the result of comparing the complete serialised text of each page with its ground truth. For newspapers this can easily mean strings of up to 20’000 words and any deviations in their order (as a result of segmentation and/or reading order detection errors) will also have an impact on this figure. This phenomenon will therefore be further explored in the next section.

#### *4.1.1.2 Strict word accuracy and document length*

An investigation into the influence of page length (and thus complexity) on word accuracy revealed a pronounced inverse correlation. At the same time, it can be observed that a decrease in word accuracy comes along with a decrease in reading order success rate (Figure 3). It seems therefore likely that reading order problems arising from the necessary text serialisation are a limiting factor for strict word evaluation.

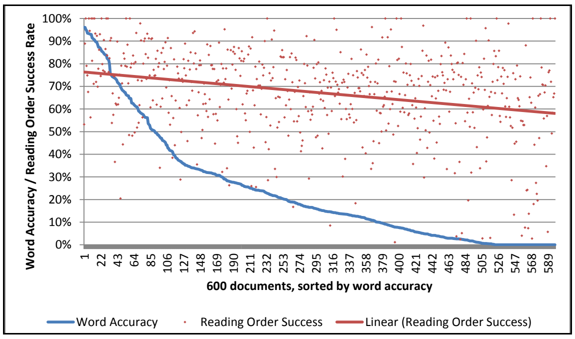

**Figure 3. Correlation plot for word accuracy and reading order success rate**

### *4.1.2 Bag of Words*

While strict word accuracy is a good measure for texts stemming from documents with a simple (linear) structure, it deviates for documents with complex layouts (such as newspapers) due to ambiguities and errors when serialising the text. To circumvent this problem it appears appropriate to carry out a Bag of Words (BoW) analysis which disregards the particular order of words.

#### *4.1.2.1 Index vs. count based*

As outlined in the metrics section, Bag of Words analysis can be done based on either an index or a count scenario. Figure 4 shows the results for both approaches (bitonal input images, FR XML result files, normalised text).

With the influence of text serialisation effects eliminated, the success rates are now much more in line with what had been expected from a manual inspection of the OCR results. From experience (discussions with library partners) it can also be said that success rates beyond 70% are usually good enough to provide an acceptable level of text search through a presentation system.

From Figure 4 it can also be observed that the index based measure is stricter than the count based one. Nevertheless, the count based measure is more likely to represent real world use scenarios than the one based on an index as it reflects not only if a document can be retrieved or not when searching for a certain term but also how it would show up in a ranked results list. The count based approach will therefore be used in the following sections on language, script, and font.

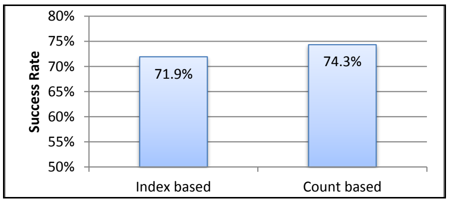

**Figure 4. Bag of Words evaluation – index vs. count based**

#### *4.1.2.2 Language*

Figure 5 shows the Bag of Words success rates for all languages (used as OCR engine parameter) in the dataset (bitonal input images, FR XML result files, normalised text, count based BoW).

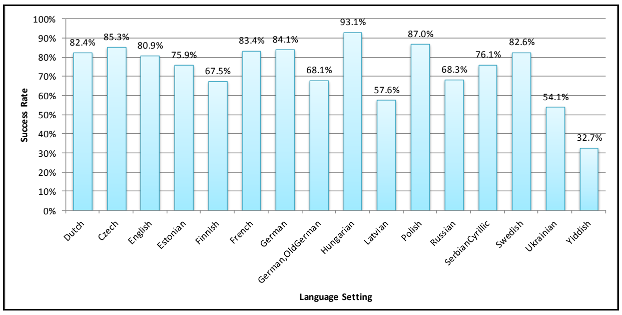

**Figure 5. Bag of Words evaluation – per language**

It can be seen that most major language are in the region of 80% and better while there is also a number of languages performing below 70%. The reason for these lower success rates may lie in the fact that languages with a smaller base of native speakers and thus documents in use are not as well supported in the OCR engine as the other languages. Another possible explanation may be the higher complexity and/or difficulty of certain scripts and languages (e.g. Old German, Yiddish).

#### *4.1.2.3 Script*

Script is an OCR setting which typically follows from the language. Figure 6 shows the performance for three different scripts that were included in the evaluation dataset (bitonal input images, FR XML result files, normalised text, count based BoW).

The main observation is that the two major scripts Latin and Cyrillic perform almost equally well. As perhaps had to be expected, less common scripts like Yiddish are not too well supported at this point. A count based Bag of Words success rate of 36.5% is usually far too low for providing text search or to display the recognised text in a presentation system. Further training and/or more specialised OCR engines would be required in order for this material to be recognisable with higher accuracy.

It has to be noted that, corresponding to the collections of the partner libraries, the sizes of the three script subsets are not equal. Nevertheless, this analysis can give a rough indication of the actual underlying trends.

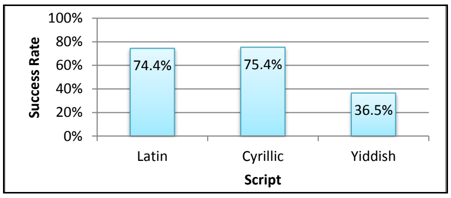

**Figure 6. Bag of Words evaluation – per script**

#### *4.1.2.4 Font*

OCR engines normally support numerous fonts without the need to specify which one(s) to expect on a page. There are, however, a few cases which are treated separately. Figure 7 shows the performance for the three font settings (Gothic, Normal, Mixed) as they were used in the production workflow (bitonal input images, FR XML result files, normalised text, count based BoW).

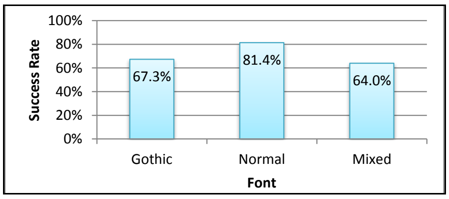

**Figure 7. Bag of Words evaluation – per font**

As was expected, normal (Antiqua) fonts are recognised best. This can be seen as a result of commercial OCR products traditionally focusing on modern business documents. However, recent developments, such as the improvement of Abbyy FineReader for Fraktur (as a result of the EC-funded project IMPACT), have led to significantly improved results for historical documents compared to what was possible a few years ago. What used to be near random results for Gothic (Fraktur) documents is now close to 70% which is considered by many the threshold for meaningful full text search. Documents with mixed content (which basically requires the OCR engine to apply all classifiers and then to decide which result to use) are still harder to recognise and this also shows that it can be very beneficial to do a proper triage in the OCR workflow and only to apply the appropriate parameters rather than letting the OCR run in auto mode.

## 4.2 Scenario-Based Evaluation

After the purely text-based assessment of OCR results in the previous section, more sophisticated aspects such as layout and reading order will now be considered.

### *4.2.1 Overall performance*

Figure 8 shows the overall performance scores for the five use scenarios that were defined above (bitonal input images, FR XML result files). Being obtained from the same actual OCR output they represent how suitable the material is for providing the respective kind of service to the end users of digital libraries. Indirectly, they do also reflect how strict the requirements are on the accuracy of the recognised material in order to implement a satisfactorily working solution.

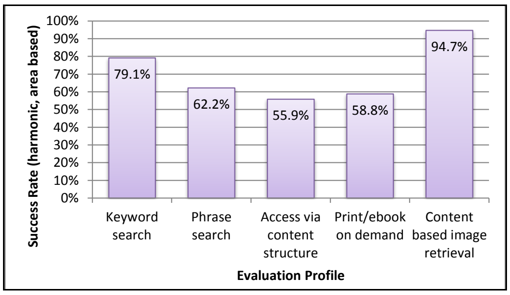

**Figure 8. Layout analysis performance per use scenario**

With an overall performance of close to 80% it can be stated that the produced material should on average be well suited for typical *Keyword search* use scenarios. The same is true for *Content based image retrieval* which has the lowest requirements, leading to the highest score. *Phrase search*, due to high requirements on segmentation and reading order, may be possible in many cases but might also lead to unsatisfactory results for newspapers with more complex layouts. *Print/ebook on demand* and *Access via content structure* come last (although not very far behind) as a result of requiring a nearly perfectly recognised layout in order to be implemented properly.

### *4.2.2 Error types*

The individual types of errors leading to the above overall scores are shown in Figure 9 and discussed in more detail below (bitonal input images, FR XML result files).

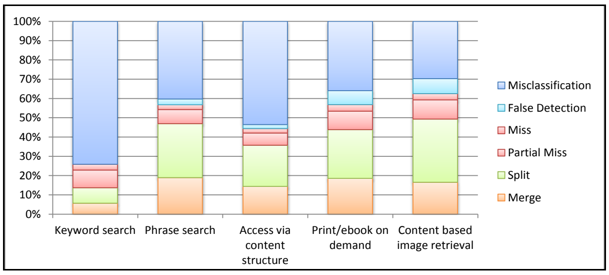

**Figure 9. Breakdown of errors**

#### *4.2.2.1 Keyword search in full text*

This use scenario is entirely centred on text regions. Since it is only important to detect all words on a page and not precise shapes and separation of regions, merge and split errors have less weight and therefore have less impact on the overall result. Misclassification and miss of text regions, on the other hand, are fatal for keyword search because crucial information is lost for succeeding processing steps. False detection is disregarded completely and does not appear in the chart at all. It can be observed that classification should be the main focus for improving the underlying analysis methods.

#### *4.2.2.2 Phrase search in full text*

In this scenario, shape and separation (segmentation) of text regions are of more importance as text phrases should not be torn apart or merged with neighbours. The evaluation profile specifies a higher weight for merge and split errors, which are therefore more pronounced in the chart. Miss and misclassification are still a major problem (about half of all errors). Better separator detection (lines and whitespace) could improve the recognition results considerably.

#### *4.2.2.3 Access via content structure*

The intention of this scenario is to extract the content structure of documents and then to allow access via linked elements (such as a table of contents linked to headings). This information is mostly encoded in regions of type heading, page number, and table of contents. Any error that compromises this information is problematic (merge of heading with main text body, misclassification as other text type, false detection of a page number, etc.). Similar to the previous scenario, merge, split, and misclassification represent the biggest part of the overall error. Multi-page recognition approaches may help detecting page numbers and running headers more reliably.

#### *4.2.2.4 Print/eBook on demand*

This slightly more generic scenario requires a profile that penalises all layout analysis errors. The main focus, however, lies on text regions (higher weights than for other types of regions). The chart shows that no individual error type can be singled out as the main problem. Due to the even distribution of error types it can only be stated that normal incremental improvements of OCR engines, especially with regard to their layout analysis capabilities, should lead to better recognition quality.

#### *4.2.2.5 Content based image retrieval*

In this final scenario, only images, graphics, and captions are of interest. The intention is that in the future users should be also presented with the means to search specifically for illustrations and graphical content in newspapers. The evaluation profile is designed to penalise miss and misclassification most, hence the impact of these error types. Nevertheless, false detection poses an issue as well. This is most likely due to misrecognised noise and clutter in the document image (remnants from the digitisation process and/or aging/preservation artefacts). Split errors have a particularly high proportion, a problem that usually arises for disjoint graphics (such as illustrations without a frame around them, charts, etc.). Potential improvements for this use scenario could go in the direction of content aware segmentation algorithms as well as smart image/graphic recognition (trying to find the meaning of the depicted objects and thus maintaining their integrity).

## 4.3 Impact of Workflow Modifications

In the last part of the results section two workflow choices are to be investigated. The first is related to an external pre-processing step for binarisation and the second is about the used OCR engine.

### *4.3.1 Binarised vs. original images*

For very practical reasons (shipping huge amounts of data to the OCR production sites) the project was faced with the question whether external binarisation (as opposed to using FineReader's built-in binarisation) at the end of each library would be an acceptable option in order to reduce the amount of data to be transferred. Since sending the original files would have caused severe production delays it was decided that this would be the preferable solution unless the recognition quality would suffer too much. A pilot experiment was carried out which projected a maximum quality loss of 1%. This was deemed acceptable and accordingly implemented in the production workflow.

Now that a larger dataset has gone through the production workflow it is possible to verify this decision. Figure 10 (FR XML result files, normalised text, count based BoW) shows a deviation of just under 1%. It can therefore be confirmed that the quality projection that was made based on the pilot experiment also holds for the representative evaluation dataset.

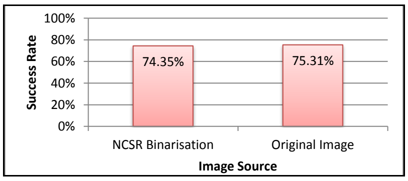

**Figure 10. External vs. internal binarisation – BoW**

Despite confirming the general decision (which was based on technical and project schedule constraints), it can be stated that using FineReader's integrated binarisation could have improved the overall Bag of Words recognition rate by about 1%.

### *4.3.2 FineReader vs. Tesseract*

FineReader was chosen as the OCR engine to be used in the Europeana Newspapers production workflow for a number of technical reasons. Being a commercial product, however, it might not always be a possible choice if license fees are an issue. In order to explore also other solutions a comparison with Tesseract, an open source OCR engine, was carried out.

#### *4.3.2.1 Text-based evaluation*

Figure 11 shows that FineReader has a considerable advantage over Tesseract in terms of text recognition (FR XML/PAGE result files, original input images, normalised text, count based BoW).

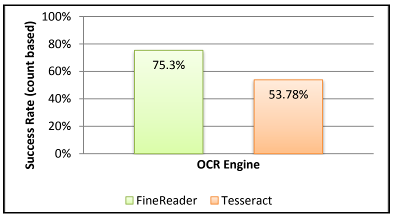

**Figure 11. FineReader Engine 11 vs. Tesseract 3.03 – BoW**

Nevertheless, Tesseract may be an interesting alternative if licensing costs are to be avoided. Moreover, Tesseract is available as source code allowing skilled developers to customise and adapt the software to specific types of documents.

#### *4.3.2.2 Scenario-based evaluation*

While Tesseract performed significantly worse than FineReader in terms of text accuracy it was surprising to see that its layout analysis capabilities are not far behind (for one use scenario Tesseract performed even better). Figure 12 shows a direct comparison of FineReader and Tesseract for the five use scenarios (FR XML/PAGE result files, original input images).

## 5. CONCLUDING REMARKS

This paper presents a detailed overview of the evaluation results which were obtained from the main Europeana Newspapers OCR production workflow based on a representative dataset collected from all partner libraries in the project.

In general it can be concluded that the produced results, especially with regard to the overall text accuracy, are of good quality and fit for use in a number of use scenarios. Moreover, technical decisions made during the setup of the production workflow could be confirmed. A number of observations (e.g. on the recognition performance for certain languages and particular layout problems) show mainly the limitations of current state-of-the-art methods rather than issues with the implemented workflow. In terms of layout analysis capabilities there is still room for improvement. Any progress in this area is expected to have a great impact on the usefulness of OCR results for more sophisticated use scenarios.

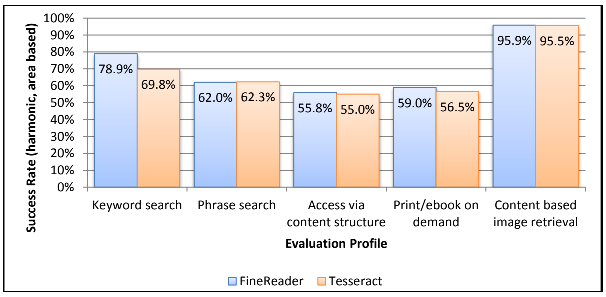

**Figure 12. FineReader Engine 11 vs. Tesseract 3.03 – Layout Analysis performance**

## REFERENCES

[1] Project ENUMERATE: http://www.enumerate.eu

[2] R. Holley, "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs", D-Lib Magazine, Volume 15, Number 3/4, 2009.

[3] S. Tanner, T. Muñoz and P.H. Ros, "Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive", D-Lib Magazine, Volume 15, Number 7/8, 2009.

[4] IMPACT - Improving Access to Text: http://www.impact-project.eu

[5] Europeana Newspapers: http://www.europeana-newspapers.eu

[6] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher , "ICDAR2013 Competition on Historical Book Recognition – HBR2013", Proc. ICDAR2013, Washington DC, USA, August 2013, pp. 1491-1495.

[7] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher , "ICDAR2013 Competition on Historical Newspaper Layout Analysis – HNLA2013", Proc. ICDAR2013, Washington DC, USA, August 2013, pp. 1486-1490.

[8] A. Antonacopoulos, C. Clausner, C. Papadopoulos and S. Pletschacher, "ICDAR2015 Competition on Recognition of Documents with Complex Layouts – RDCL2015", Proc. ICDAR2015, Tunisia, August 2015.

[9] The European Library: http://www.theeuropeanlibrary.org

[10] Europeana: http://www.europeana.eu

[11] ABBYY FineReader SDK: http://www.abbyy.com/ocr-sdk/

[12] ALTO: Analyzed Layout and Text Object:
http://www.loc.gov/standards/alto/

[13] METS: Metadata Encoding & Transmission Standard,
http://www.loc.gov/standards/mets/

[14] S. V. Rice, "Measuring the Accuracy of Page-Reading Systems", Dissertation, University of Nevada, 1996.

[15] C. Clausner, S. Pletschacher, A. Antonacopoulos, "Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods", Proc. ICDAR2011, Beijing, China, September 2011, pp. 1404-1408.

[16] C. Clausner, S. Pletschacher, A. Antonacopoulos, "The Significance of Reading Order in Document Recognition and its Evaluation", Proc. ICDAR2013, Washington DC, USA, August 2013, pp. 688-692.

[17] C. Clausner, C. Papadopoulos, S. Pletschacher and A. Antonacopoulos, "The ENP Image and Ground Truth Dataset of Historical Newspapers", Proc. ICDAR2015, Tunisia, August 2015.

[18] http://www.primaresearch.org/tools

[19] S. Pletschacher and A. Antonacopoulos, "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework", Proc. ICPR2008, Istanbul, Turkey, August 2010, pp. 257-260.