DELIVERABLE

Project Acronym: Europeana Newspapers

Grant Agreement number: 297380

Project Title: A Gateway to European Newspapers Online

D3.5 Performance Evaluation Report

Revision: 1.0

Authors: **Stefan Pletschacher, USAL**
Christian Clausner, USAL
Apostolos Antonacopoulos, USAL

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
PU	**Public**	
PP	**Restricted to other programme participants (including Commission Services)**	**x**
RE	**Restricted to a group specified by the consortium (including the Commission Services)**	
CO	**Confidential, only for members of the consortium and the Commission Services**	

Revision History

Revision	Date	Author	Organisation	Description
0.1	07-07-2014	Christian Clausner Stefan Pletschacher	USAL	First draft
0.2	21-07-2014	Stefan Pletschacher	USAL	Updated version
0.3	29-07-2014	Christian Clausner Stefan Pletschacher	USAL	Updated version
0.4	30-07-2014	Stefan Pletschacher Christian Clausner Apostolos Antonacopoulos	USAL	Final version for internal review
0.5	31-07-2014	Stefan Pletschacher	USAL	Minor corrections
0.6	31-07-2014	Clemens Neudecker	SBB	Internal review
1.0	31-07-2014	Sandra Kobel	SBB	Final version

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of Contents

1. Introduction

This report 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 newspaper pages (for which ground truth was created within the scope of *T3.2 Evaluation datasets*) 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 (as defined in *D3.1 Evaluation profiles for use scenarios*). Moreover, this report confirms the general success of the refinement process that was implemented in Work Package 2 and the validity of workflow-related decisions that were made based on experiments and feedback within the scope of *T3.5 Impact of refinement strategies*.

2. 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 on the layout being 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 that now wrongly goes between 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 partners and material in Europeana Newspapers, a survey was carried out within the scope of *T3.1 Use scenarios* which resulted 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 following evaluation profiles which represent settings and error weights corresponding to the five use scenarios (following *D3.1 Evaluation profiles for use scenarios*):

2.1 Keyword search in full text

Summary:

- Only text regions are of interest
- Miss of regions or parts of regions is penalised most
- Splits are penalised only a little (a keyword may have been split)
- Merges are less important (only merges across columns may be problematic when hyphenation is involved)
- Misclassification from text to text is irrelevant (e.g. paragraph misclassified as heading)
- False detection is irrelevant (additional regions are unlikely to compromise the indexing)
- Reading order is ignored (only the occurrence of words is of interest, no matter in which order)
- Bag of Words evaluation for text is sufficient

2.2 Phrase search in full text

Summary:

- Only text regions are of interest
- Miss of regions or parts of regions is penalised most
- Merge of regions not in reading order is highly penalised
- Merge or split of consecutive text blocks ('allowable') only minimally penalised
- Splits are unwanted but not especially emphasized (default penalty)
- False detections are disregarded
- Focus on word accuracy for text evaluation (high accuracy required)

2.3 Access via content structure

Summary:

- Focus on textual elements
- Special emphasis on subtypes headings, page numbers and TOC-entries
- Miss, partial miss and misclassification is penalised most
- Merge and split of regions not in reading order is highly penalised
- Merge and split of consecutive text blocks ('allowable') only minimally penalised
- False detection penalised least
- Reading order important
- Focus on word accuracy (moderate to high requirements on text accuracy)

2.4 Print/eBook on demand

Summary:

- Text regions are considered more important than other regions
- Miss of regions or parts of regions is penalised most
- Merges get a high penalty
- Merge or split of consecutive text blocks ('allowable') only minimally penalised
- Image, graphic and line drawing are treated as equal (misclassification not penalised)
- Noise and unknown regions are irrelevant
- Reading order important
- Focus on word accuracy (moderate to high requirements on text accuracy)

2.5 Content based image retrieval

Summary:

- Only image, graphic, line drawing and chart are of interest
- Image, graphics, line drawing and chart are considered one class (no misclassification error)
- Miss, partial miss and misclassification are penalised most
- Reading order and allowable splits, merges are disregarded
- Low requirements on text accuracy (only captions)

3. Metrics

Each scenario defines an evaluation strategy that includes settings and weights which are then to be 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.

3.1 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. This measure is described in S. V. Rice's dissertation 'Measuring the accuracy of page-reading systems'[1]. 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.

Rice also describes a *character accuracy* measure which uses the same principles as the word accuracy only that, instead of edits, deletes, and inserts of whole words, the character level is used. Due to the nature of the algorithm, however, 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 typical use scenarios.

The need to handle text serialisations of potentially very long documents, as is typically 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 if 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).

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

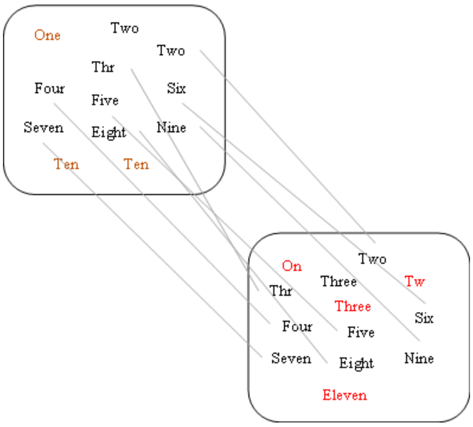


Figure 1: Schematic depiction of the Bag of Words text evaluation approach

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

3.2 Layout-based evaluation

In addition to textual results, page reading systems (such as Abbyy FineReader) 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).

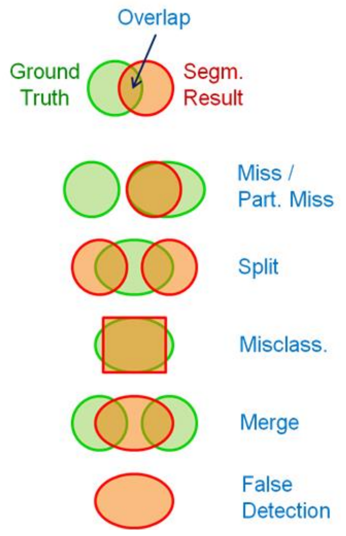
Which of those measures is to be used and how much impact they should have on the overall result is specified in evaluation profiles. These include weights for segmentation errors (merge, split, miss, and false detection), misclassification errors, and reading order errors. Depending on the profile, the overall success rate for an OCR result can vary significantly.

The next subsections detail the specific evaluation methods.

3.2.1 Evaluation of segmentation and classification results

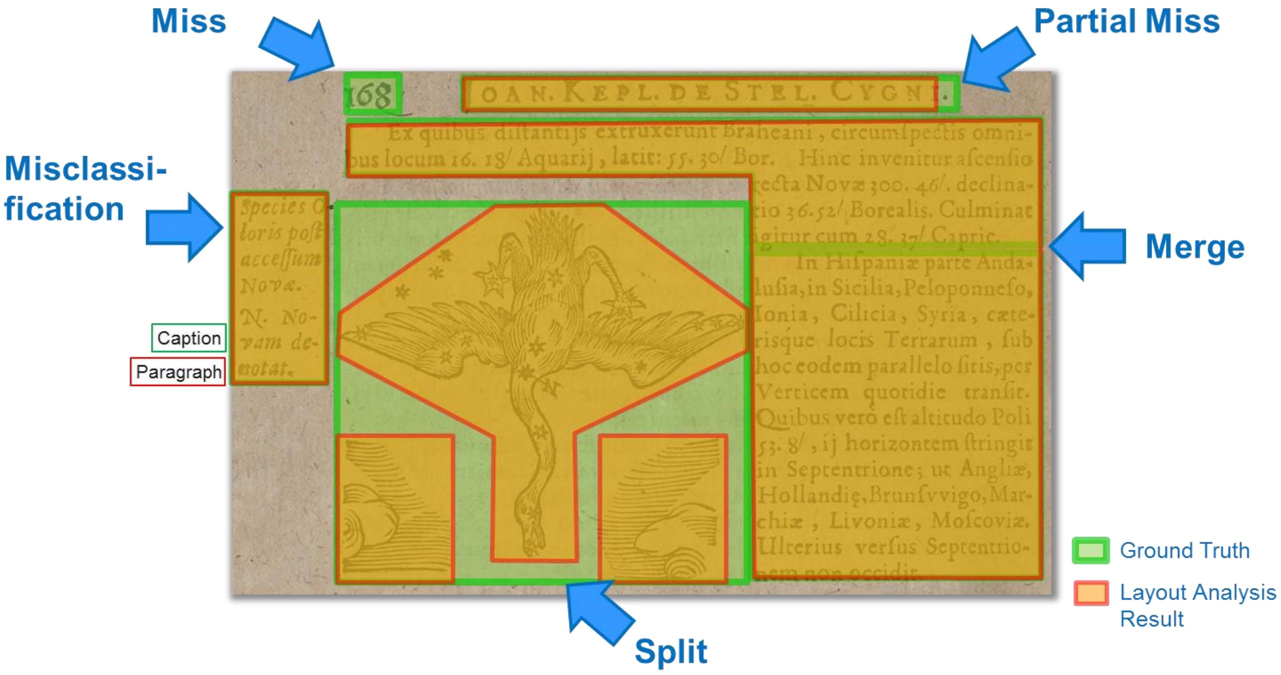
The evaluation algorithm[2] takes into account a wide range of situations and provides considerable details on performance of layout analysis methods. The system performs a geometric comparison between regions detected by a segmentation method and ground-truth regions in order to identify erroneously merged, split, missed, partially missed or misclassified regions. Each type of error is weighted according to the types of regions involved and the situation they are found in. Figure 2 provides an overview of the different error types and Figure 3 shows how an OCR result and its corresponding ground truth are compared in order to ascertain potential errors.

[2] Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods, C. Clausner, S. Pletschacher, A. Antonacopoulos, Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR2011), Beijing, China, September 2011, pp. 1404-1408.



- Miss/partial miss: A ground truth region is not or not completely overlapped by a segmentation result region.
- Split: A ground truth region is overlapped by more than one segmentation result region.
- Misclassification: A ground truth region is overlapped by a segmentation result region of another type.
- Merge: A segmentation result region overlaps more than one ground truth region.
- False detection: A segmentation   result region overlaps no ground truth region.

Figure 2: Layout evaluation error types



168 IOAN. KEPL. DE STEL. CYGNI.

Ex quibus diſtantijs extruxerunt Braheani, circumſpectis omnibus locum 16. 18/ Aquarij, latit: 55. 30/ Bor. Hinc invenitur aſcenſio recta Novæ 300. 46/. declinatio 36.52/ Borealis. Culminat igitur cum 28. 37/ Capric.

In Hiſpaniæ parte Andaluſia, in Sicilia, Peloponneſo, Ionia, Cilicia, Syria, cæteriſque locis Terrarum, ſub hoc eodem parallelo ſitis, per Verticem quotidie tranſit. Quibus verò eſt altitudo Poli 53. 8/, ij horizontem ſtringit in Septentrione; ut Angliæ, Hollandię, Brunſvvigo, Marchiæ, Livoniæ, Moſcoviæ. Ulterius verſus Septentrionem non occidit.

Species Coloris poſt acceſſum Novæ.

N. Novam denotat.

Figure 3: Layout evaluation - comparison of result and ground truth

3.2.2 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. 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. Figure 4 shows an example of a document page which includes groups of ordered and unordered content elements.

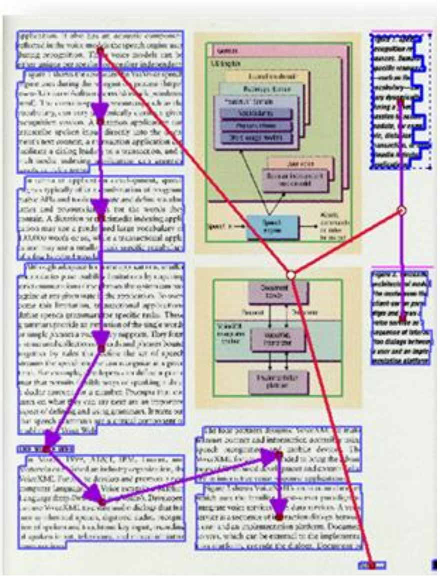
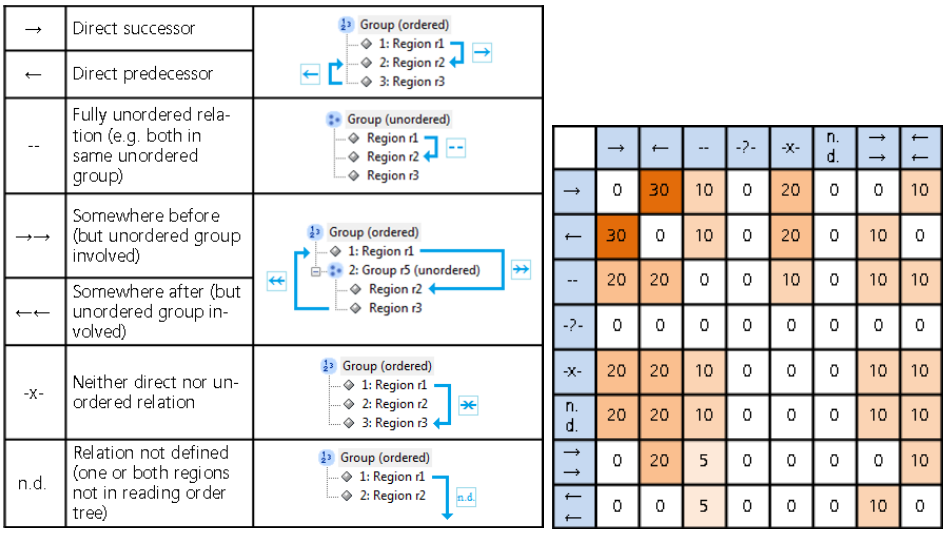


Figure 4: Reading order involving groups of ordered and unordered elements.

The method[3] 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 penalised according to below matrix and are finally combined to a composite penalty.

[3] The Significance of Reading Order in Document Recognition and its Evaluation, C. Clausner, S. Pletschacher, A. Antonacopoulos, Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR2013), Washington DC, USA, August 2013, pp. 688-692.



→	Direct successor	Group (ordered) 1: Region r1 2: Region r2 3: Region r3 → ←
←	Direct predecessor	
--	Fully unordered relation (e.g. both in same unordered group)	Group (unordered) Region r1 Region r2 Region r3 --
→→	Somewhere before (but unordered group involved)	Group (ordered) 1: Region r1 2: Group r5 (unordered) Region r2 Region r3 →→ ←←
←←	Somewhere after (but unordered group involved)	
-x-	Neither direct nor unordered relation	Group (ordered) 1: Region r1 2: Region r2 3: Region r3 -x-
n.d.	Relation not defined (one or both regions not in reading order tree)	Group (ordered) 1: Region r1 2: Region r2 n.d.

	→	←	--	-?-	-x-	n. d.	→→	←←
→	0	30	10	0	20	0	0	10
←	30	0	10	0	20	0	10	0
--	20	20	0	0	10	0	10	10
-?-	0	0	0	0	0	0	0	0
-x-	20	20	10	0	0	0	10	10
n. d.	20	20	10	0	0	0	10	10
→→	0	20	5	0	0	0	0	10
←←	0	0	5	0	0	0	10	0

Figure 5: Left – Possible relations between regions; Right – Penalty matrix for wrong relations.

4. 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. Task 3.2 *Evaluation datasets* was therefore focusing on the creation of a high quality dataset including ground truth files in PAGE (Page Analysis and Ground truth Elements) format. Major requirements on the dataset were for it to be:

- "Realistic – reflecting the actual library holdings with regard to representativeness and frequency of documents
- Comprehensive – including metadata and detailed ground truth
- Flexibly structured – supporting all stakeholders to search, browse, group etc. and allowing other technical systems (such as workflow systems and evaluation tools) to interface directly."[4]

4.1 Dataset creation

The dataset was created in three main stages (following D3.2 *Evaluation Dataset Including Ground Truth*):

- 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:

[4] The IMPACT Dataset of Historical Document Images, C. Papadopoulos, S. Pletschacher, C. Clausner, A. Antonacopoulos, Proceedings of the 2013 Workshop on Historical Document Imaging and Processing (HIP2013), Washington DC, USA, August 2013, pp. 123-130.

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.

4.2 Dataset statistics

In total the dataset used for evaluating the main production workflow comprised 600 newspaper pages. The following charts show an overview of the distribution of languages, scripts, and fonts (being the main parameters of OCR engines).

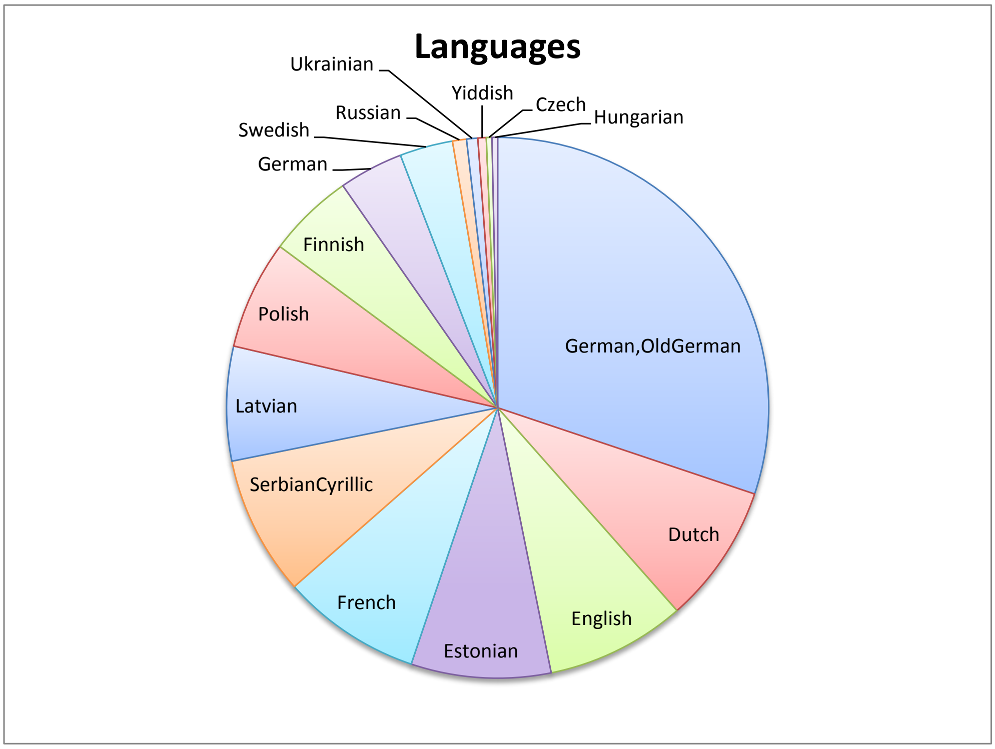


Figure 6: Distribution of languages in the evaluation dataset

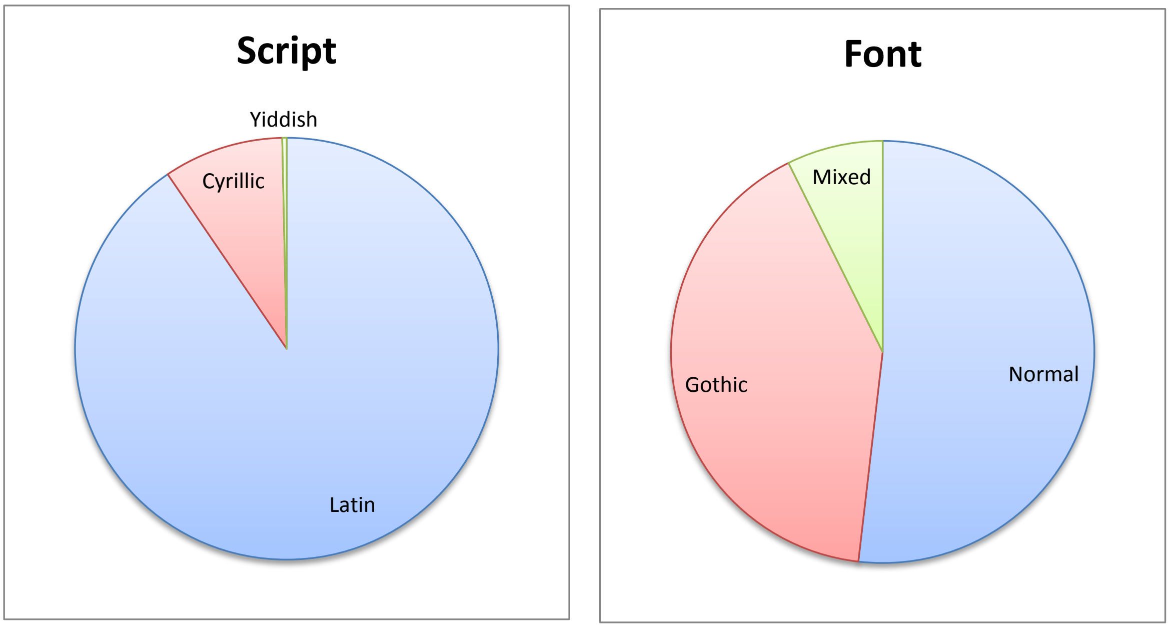


Figure 7: Distribution of scripts and fonts in the evaluation dataset

5. Evaluation Workflow

In order for the evaluation results to be objective and reproducible as well as the overall process to run as efficient as possible, an automated evaluation workflow was set up, using the tools specifically developed for this purpose within the scope of Task 3.3 *Evaluation tools*. In the following it is described how the required data was produced and processed. Figure 8 shows an overview of the overall evaluation workflow.

5.1 Ground truth production

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

5.2 OCR result production

OCR output was produced using the Europeana Newspapers production workflow which included the NCSR 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 (using the USAL PAGE Converter tool) to be used by the evaluation tools.

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

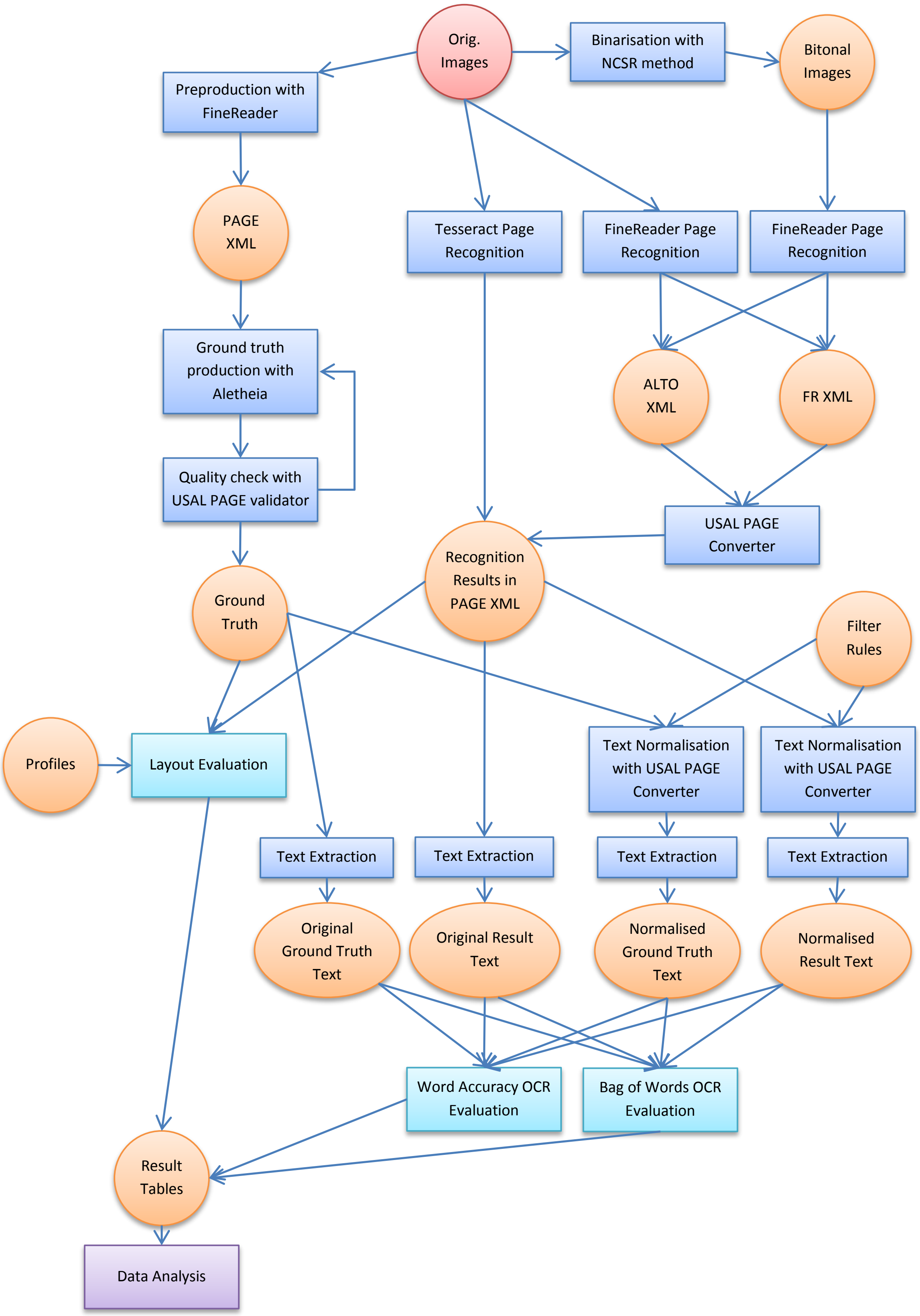

Orig. Images
Binarisation with NCSR method
Bitonal Images
Preproduction with FineReader
PAGE XML
Tesseract Page Recognition
FineReader Page Recognition
FineReader Page Recognition
Ground truth production with Aletheia
ALTO XML
FR XML
Quality check with USAL PAGE validator
USAL PAGE Converter
Recognition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

Figure 8: Overall evaluation workflow

5.3 Text-based performance evaluation

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

5.3.1 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;
- 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).

5.3.2 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, the USAL 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.

5.3.3 Evaluation

The actual performance evaluation was carried out using the USAL Text Evaluation tool 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

5.4 Layout-based performance evaluation

The evaluation was carried out using the USAL Layout Evaluation tool. Several factors were taken into account, leading to a total of 20 result tables:

- Five different evaluation profiles matching the use scenarios defined in D3.1
- OCR results based on bitonal and original images
- ALTO XML and FineReader XML format

6. Results and Discussion

This section summarises all the results that were obtained from the evaluation experiments as outlined in the previous sections. 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.

6.1 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.

6.1.1 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.

6.1.1.1 Overall results

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

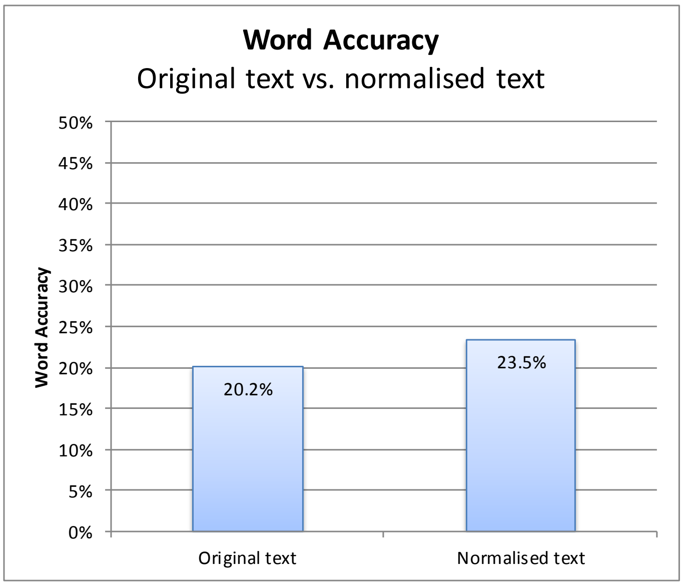


Figure 9: Overall word accuracy – original vs. normalised text

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, some 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).

The second observation is, even if looking at the relaxed measure based on normalised text, that the overall results are rather low. 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.

6.1.1.2 Strict word accuracy and document length

In order to investigate the influence of document length (and thus the complexity of a newspaper page) on the word accuracy, a comparison of six document-length classes was carried out (based on bitonal input images, ALTO result files, and normalised text).

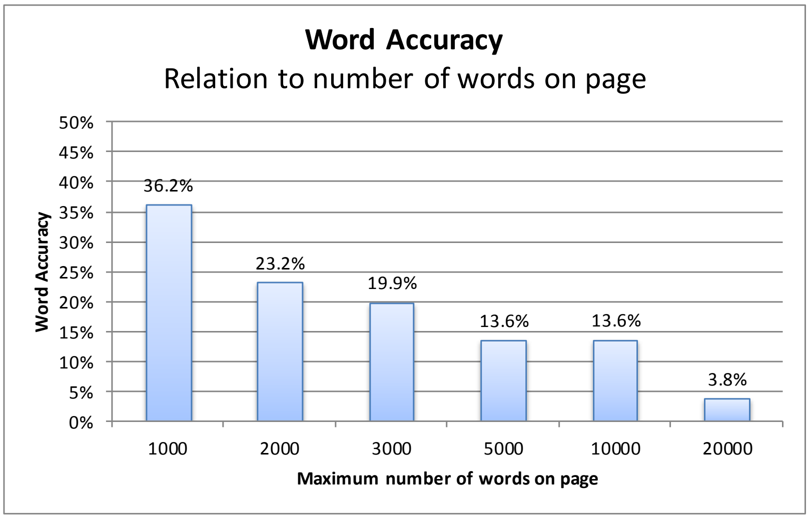


Figure 10: Word accuracy for different document-length classes

From Figure 10 it becomes immediately obvious that there is a strong inverse correlation between the length (and thus complexity) of documents and the achievable word accuracy. A correlation plot between word accuracy and reading order success rate substantiates this hypothesis (Figure 11). It can therefore be concluded that reading order problems arising from the necessary text serialisation are a limiting factor for strict word evaluation.

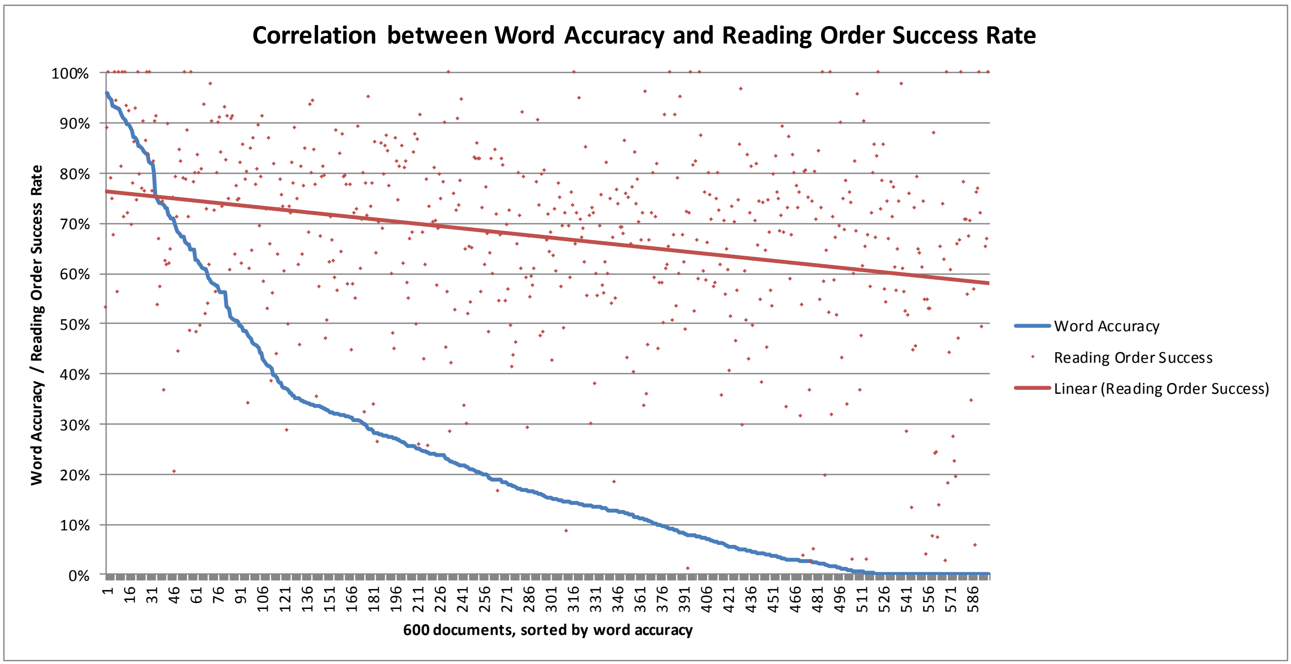


Figure 11: Correlation plot for word accuracy and reading order success rate

6.1.2 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 analysis which disregards the particular order of words.

6.1.2.1 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 12 shows the results for both approaches (bitonal input images, FR XML result files, normalised text).

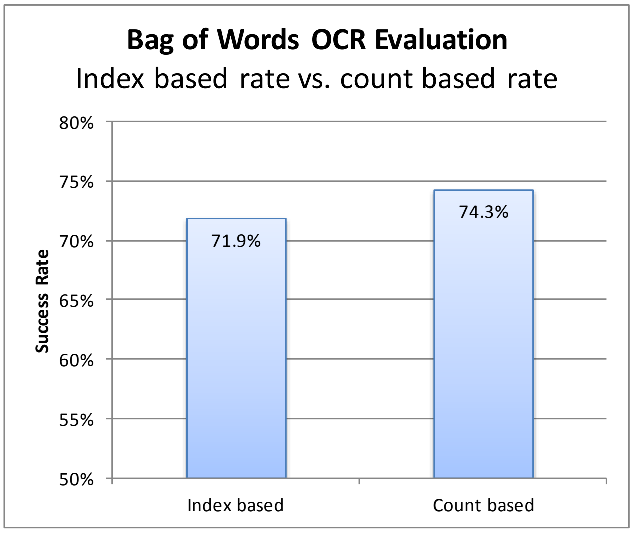


Figure 12: Bag of Words evaluation – index vs. count based

Now, with the influence of text serialisation effects eliminated, the success rates are much more in line with what had been expected from a manual inspection of the OCR results. From experience 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 12 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 ranked results.

6.1.2.2 Count based

In the following, using the count based Bag of Words evaluation approach, the effects of specific document characteristics such as language, script, and font are to be further examined.

6.1.2.2.1 Language

Figure 13 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).

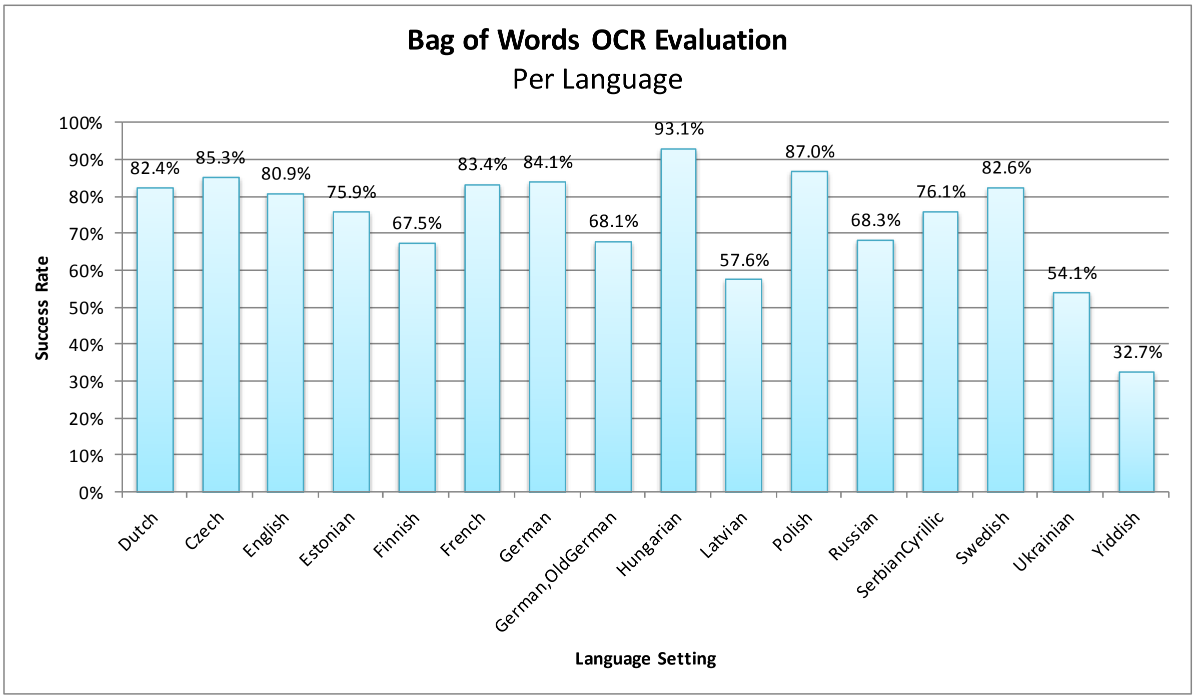


Figure 13: Bag of Words evaluation – per language

It can be seen that most major languages are in the region of 80% and better while there is also a number of languages performing below 70%. The reason for these lower results 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).

6.1.2.2.2 Script

Script is an OCR setting which typically follows from the language. Figure 14 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).

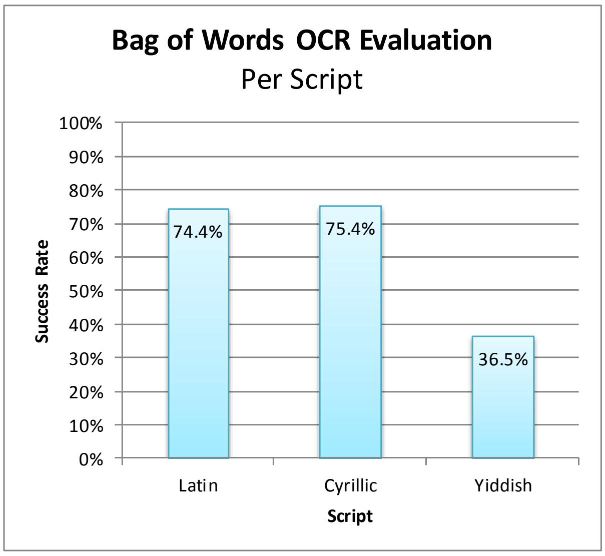


Figure 14: Bag of Words evaluation – per script

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 number of documents in the three categories is not the same. Nevertheless, this analysis can give a rough indication of the actual underlying trends.

6.1.2.2.3 Font

OCR engines normally support numerous fonts without the need to specify which one(s) to expect in the input image. There are, however, a few cases which are treated separately. Figure 15 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).

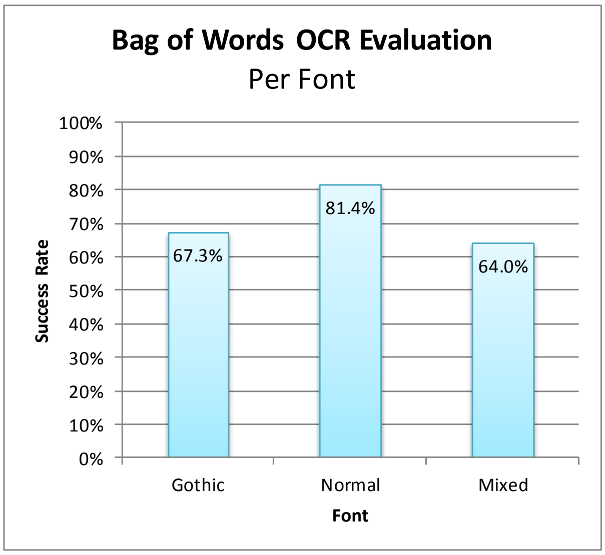


Figure 15: Bag of Words evaluation – per font

As had to be 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.

6.2 Scenario-based evaluation

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

6.2.1 Overall performance

The following chart shows the overall performance scores (as described in 3.2) for the five use scenarios that were defined in D3.1 (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.

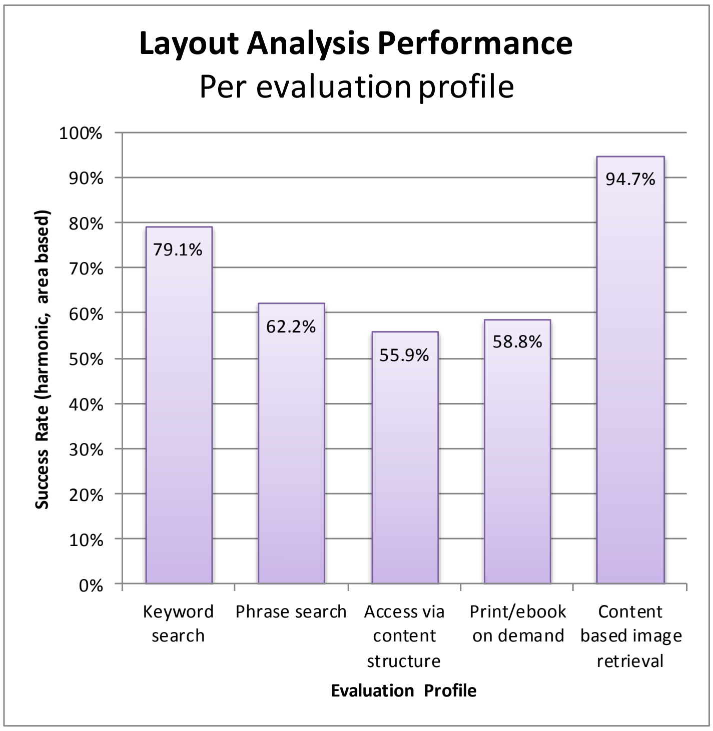


Figure 16: Layout analysis performance for different use scenario

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.

6.2.2 Error types

The individual types of errors leading to the above overall scores are discussed in more detailed in the following.

6.2.2.1 Scenario 1: Keyword search in full text

This use scenario is centred on text regions only. 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 entirely and does not appear in the chart at all (bitonal input images, FR XML result files).

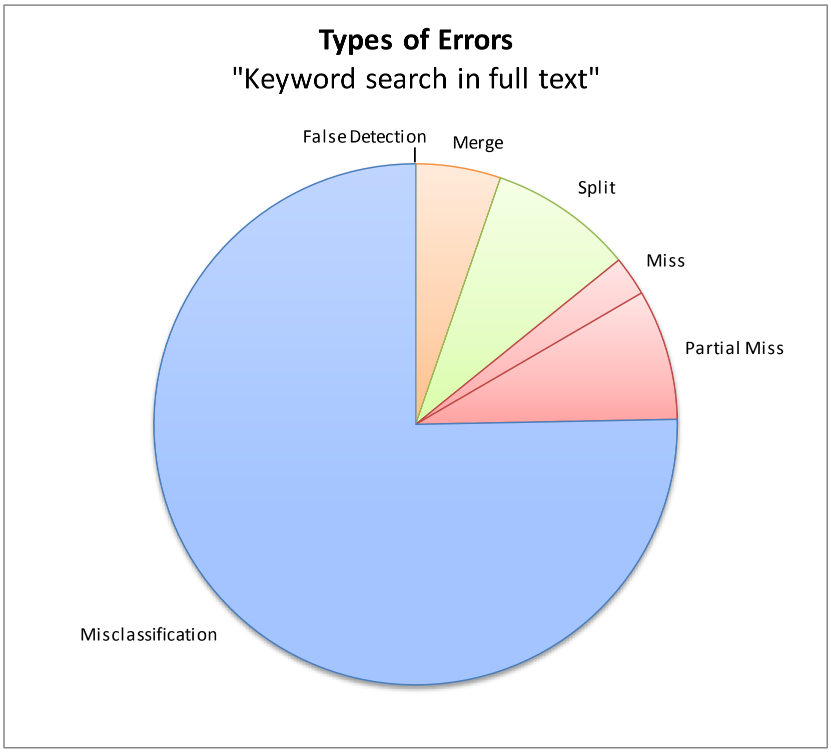


Figure 17: Proportions of layout analysis errors – Keyword search in full text

It can be observed that improving classification should be the main focus for future work.

6.2.2.2 Scenario 2: Phrase search in full text

In contrast to keyword search, shape and separation (segmentation) of text regions are of more importance in this scenario because 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 (Figure 18, bitonal input images, FR XML result files). Miss and misclassification are still a major problem (about half of the total errors).

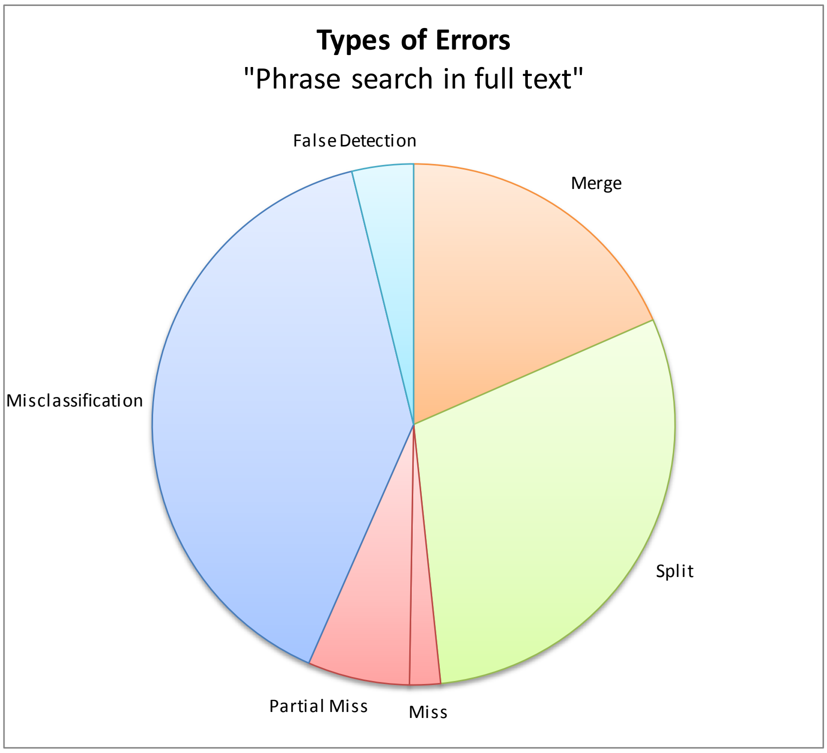


Figure 18: Proportions of layout analysis errors – Phrase search in full text

Better separator detection (lines and whitespace) could improve the recognition results considerably.

6.2.2.3 Scenario 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 (Figure 19, bitonal input images, FR XML result files).

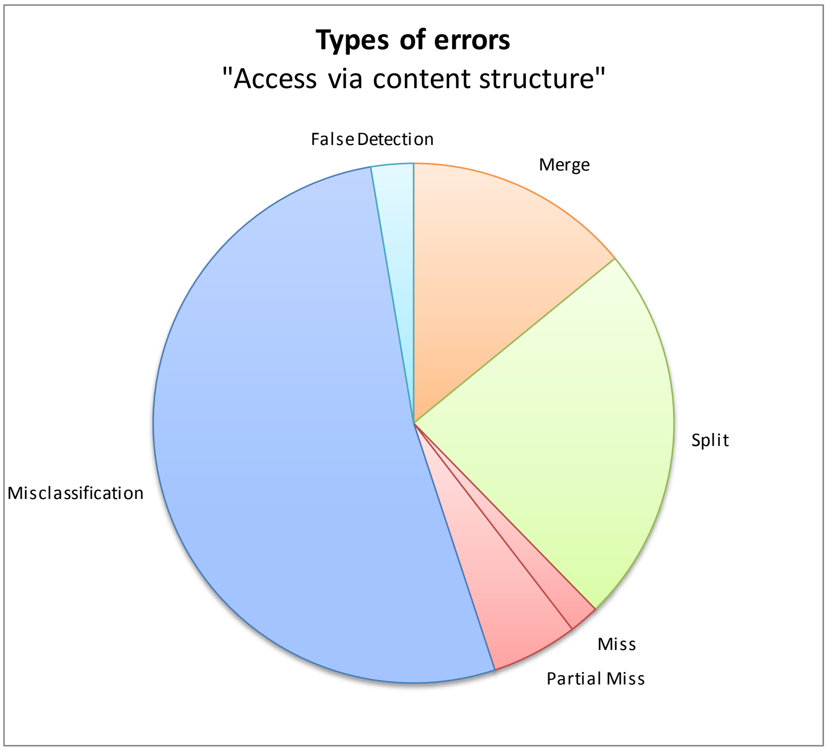


Figure 19: Proportions of layout analysis errors – Access via content structure

Multi-page recognition approaches may help detecting page numbers and running headers more reliably.

6.2.2.4 Scenario 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 (Figure 20, bitonal input images, FR XML result files).

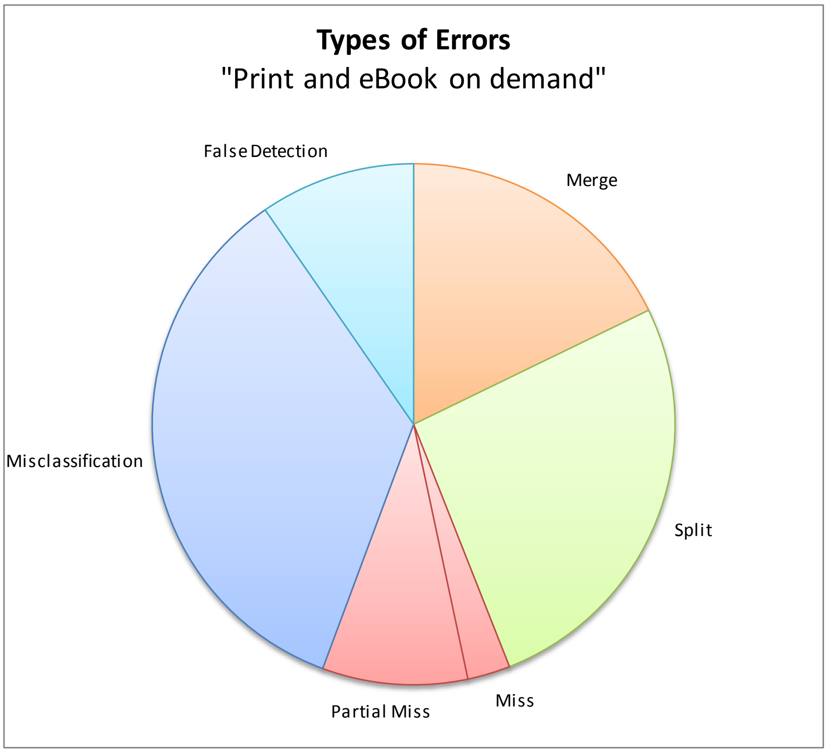


Figure 20: Proportions of layout analysis errors – Print/eBook on demand

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 raise the recognition quality in the future.

6.2.2.5 Scenario 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. Nevertheless, false detection poses an issue as well. This is most likely due to misrecognised noise and clutter on the document image (remnants from the digitisation process and/or aging/preservation artefacts). Split errors have a particularly high proportion (Figure 21, bitonal input images, FR XML result files), a problem that usually arises for disjoint graphics (such as illustrations without a frame around them, charts, etc.).

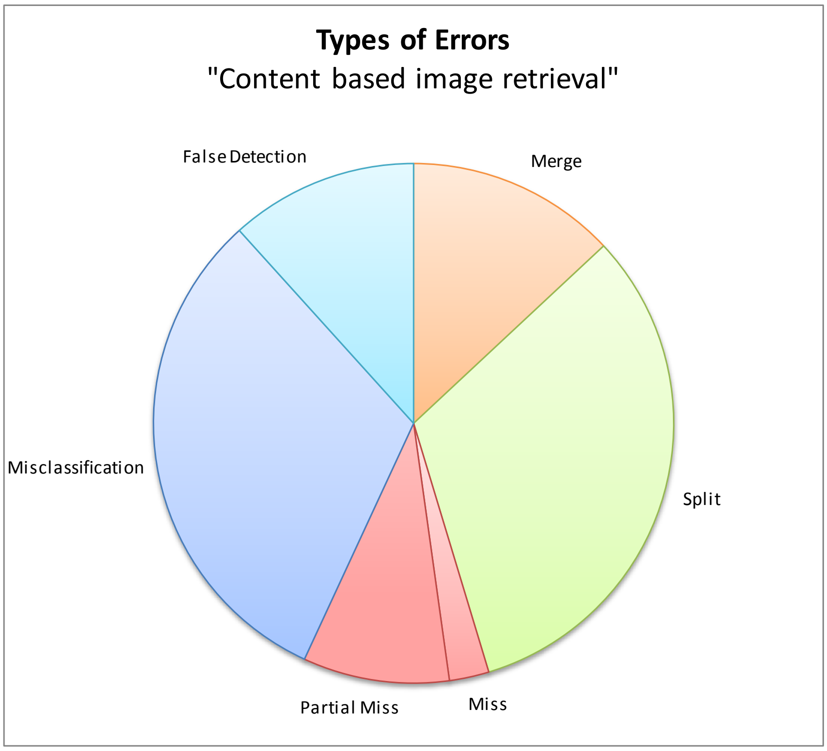


Figure 21: Proportions of layout analysis errors – Content-based image retrieval

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).

6.3 Impact of workflow modifications

In the last part of the Results and Discussion 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.

6.3.1 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 (based on a small dataset) was carried out within Work Package 3 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 time to verify this decision. Figure 22 (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.

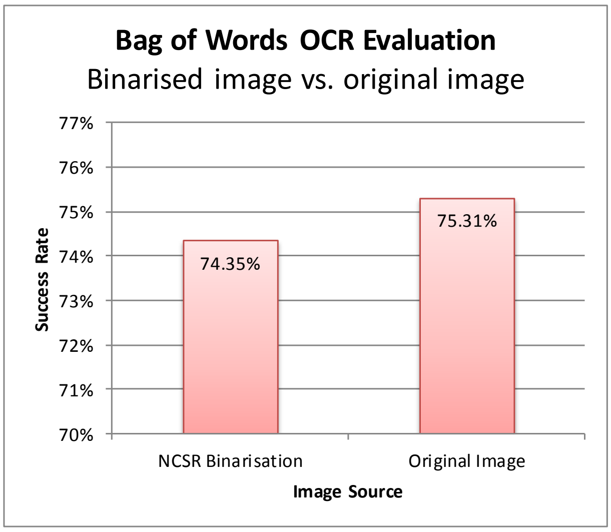


Figure 22: OCR results for external and internal binarisation

Despite confirming the general decision (which was based on a technical/scheduling necessity), it can be stated that using FineReader's integrated binarisation could have improved the overall Bag of words recognition rate by about 1%.

6.3.2 FineReader vs. Tesseract

FineReader was chosen as the OCR engine to be used in the Europeana Newspapers production workflow for numerous 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.

6.3.2.1 Text-based evaluation

Figure 23 shows that FineReader as the professional solution has a considerable advantage over Tesseract when it comes to pure text recognition (FR XML/PAGE result files, original input images, normalised text, count based BoW). When off-the-shelf software is required it is therefore the recommended solution. Nevertheless, Tesseract may be an interesting alternative if license fees 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.

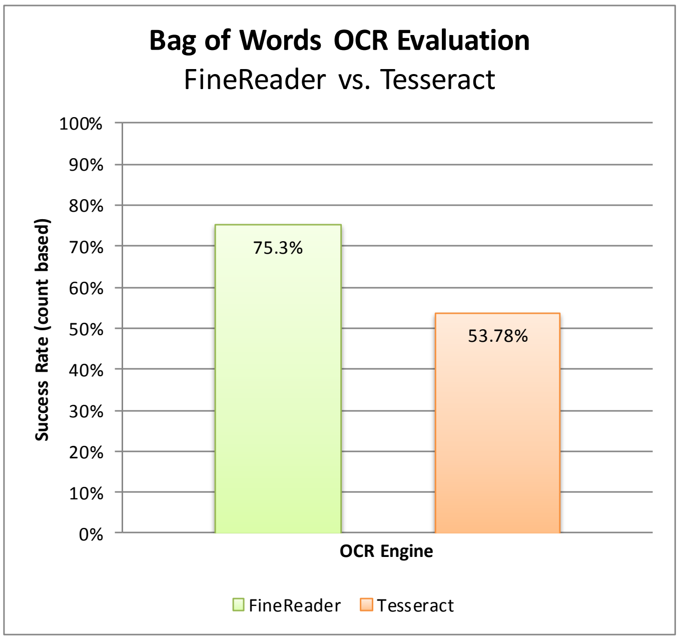


Figure 23: FineReader Engine 11 vs. Tesseract 3.03 – Bag of Words

6.3.2.2 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 24 shows a direct comparison of FineReader and Tesseract for the five use scenarios from before (FR XML/PAGE result files, original input images).

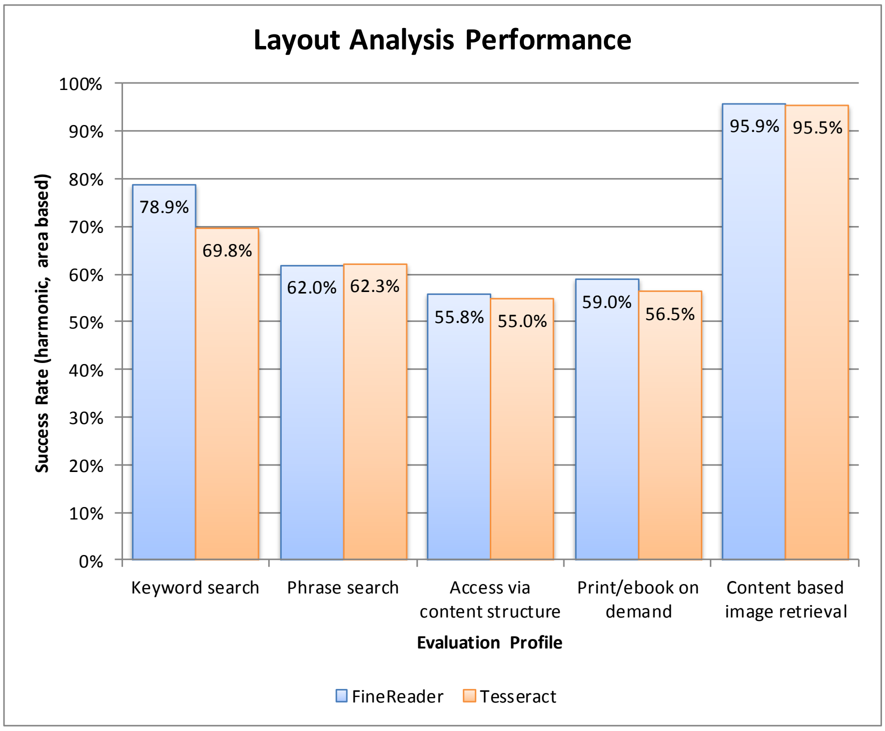


Figure 24: FineReader Engine 11 vs. Tesseract 3.03 – layout analysis performance

Overall, FineReader still comes out as the better choice though but Tesseract has a lot of potential and the fact that it can be used for free playing in its favour.

7. Additional Observations

Recognition results from both Abbyy FineReader and Tesseract OCR suffered from some issues which needed correction/post-processing:

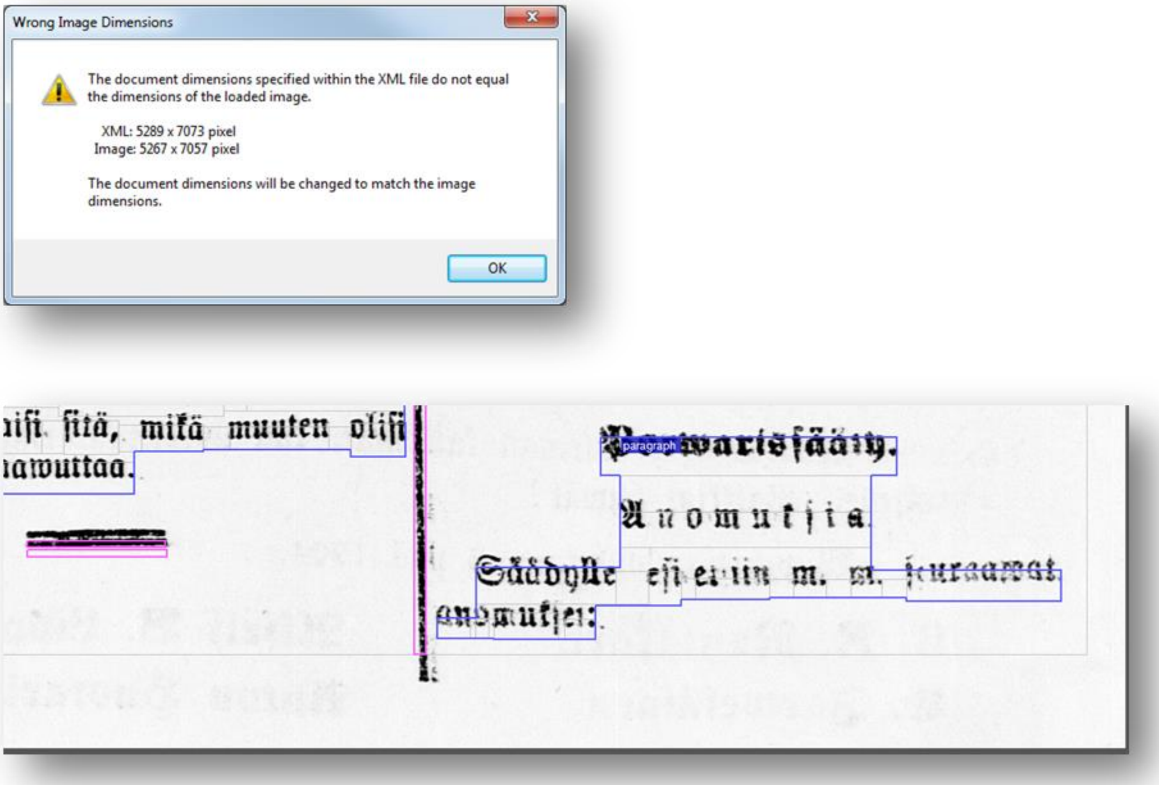
7.1 Negative coordinates

Both recognition engines in some cases did output regions with negative coordinates. Even though in most cases this was limited to “-1”, negative coordinates are invalid and needed to be corrected.

7.2 Differing dimensions for image and OCR result

FineReader OCR results (ALTO or FineReader XML) sometimes contain contradictory page dimensions in comparison to the corresponding image file (e.g. page dimensions in XML file bigger than the actual image dimensions). This might be due to enabled skew correction in FineReader which (internally) creates a larger image that is not available as output.

As a solution the dimensions specified in the XML file were set to the image dimensions. Since the difference was minor in most cases, this can be seen as an acceptable inaccuracy. Nevertheless, in some use scenarios (e.g. print/eBook on demand) this could lead to noticeable problems.



Wrong Image Dimensions

The document dimensions specified within the XML file do not equal the dimensions of the loaded image.

XML: 5289 x 7073 pixel
Image: 5267 x 7057 pixel

The document dimensions will be changed to match the image dimensions.

Figure 25: Impact of deviating coordinates in original image and OCR result

7.3 No explicit reading order

Abbyy FineReader does not explicitly specify the order of text regions, even though the ALTO format contains an appropriate mechanism (“IDNEXT”). Since the reading order is required for several evaluation approaches, it was created from the sequence of the regions as they appear in the XML result files. This approach leads to acceptable and generally meaningful results.

7.4 Recognition failures

Tesseract completely failed to recognise a small number of newspaper pages due to internal errors. This underlines the difference between a robust commercial product like Abbyy FineReader and an open source system like Tesseract. Missing results were treated as zero success in the evaluation (by creating empty result files).

8. Conclusion

This report presented 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 that were 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 and any progress in this area could have a great impact on the usefulness of OCR results for more sophisticated use scenarios.

While this report officially completes Task 3.5 *Impact of refinement strategies,* its findings and lessons learned will be further used within the scope of Task 3.6 *Planning resources and quality estimation tools* which runs until the end of the project.