Making Europe’s Historical Newspapers Searchable[†]

Clemens Neudecker
Directorate General
Staatsbibliothek zu Berlin
Preußischer Kulturbesitz
Berlin, Germany

Apostolos Antonacopoulos
Pattern Recognition & Image Analysis Research Lab
School of Computing, Science and Engineering
University of Salford
Greater Manchester, United Kingdom

Abstract—This paper provides a rare glimpse into the overall approach for the refinement, i.e. the enrichment of scanned historical newspapers with text and layout recognition, in the Europeana Newspapers project. Within three years, the project processed more than 10 million pages of historical newspapers from 12 national and major libraries to produce the largest open access and fully searchable text collection of digital historical newspapers in Europe. In this, a wide variety of legal, logistical, technical and other challenges were encountered. After introducing the background issues in newspaper digitization in Europe, the paper discusses the technical aspects of refinement in greater detail. It explains what decisions were taken in the design of the large-scale processing workflow to address these challenges, what were the results produced and what were identified as best practices.

Keywords—historical newspapers; optical character recognition; layout analysis; digital libraries

I. Introduction

Newspapers are amongst the most valuable sources for scholars that are interested in researching public opinion and how it has been shaped over time. Moreover, Europe’s historical newspapers together can provide a rich resource on how developments and ideas were perceived and spread across different countries. However, due to many technical challenges, the large-scale processing of historical newspapers with text and layout recognition technologies has long been lacking. The Europeana Newspapers project [1] addressed this by aggregating and refining (adding text and layout information) more than 10 million historical newspaper pages in order to make them fully searchable.

There has been an increasing number of digitisation projects in libraries, yet only a fraction of actual document holdings have been digitized (only 4% in national libraries, for instance) [2]. Moreover, most of that digitized content is in the form of scanned pages with no associated text or other refinement. This presents a significant opportunity for document analysis researchers and system developers. However, very little is usually known about the real world issues (technical and others) surrounding such digitisation projects, as they are often commissioned to commercial service providers. Those issues and characteristics play a significant role in determining the direction and priorities of real-world document analysis systems research and development.

This paper presents valuable insights into the nature of and experiences gained from the project which created the largest open access and fully searchable text collection of digitized historical newspapers in Europe. It provides the context and technical approach to addressing the diversity of material, dispersed content, breadth of quality and access regulations, significantly amplified by the very large scale of the overall undertaking. The authors hope that the presented knowledge gained through experience and the information on real-world decisions and their outcomes will be valuable for anyone developing or configuring document analysis systems.

A brief description of the project is given in the next section. In Section III, the background on the selection of the material to be processed, based on the challenges of its nature, accessibility, legal considerations and user requirements is presented. Section IV describes and discusses the different aspects of the refinement process and related issues in detail. The paper ends with concluding remarks made in Section V.

II. The Europeana Newspapers Project

Europeana Newspapers [1] was a 38-month Best-Practice-Network funded by the European Commission under the theme CIP-ICT-PSP.2011.2.1 - Aggregating content for Europeana. The project ran from February 1st 2012 to March 31st 2015 and included 18 project partners (the majority of which were national or major European Libraries), 11 associated partners and 35 networking partners. It was set up to accomplish, among others, the following ambitious targets:

- Make historical newspapers fully searchable by refining 10 million scanned pages with Optical Character Recognition (OCR) and Optical Layout Recognition (OLR) and experimental support for Named Entity Recognition (NER) in three languages.
- Make digital historical newspapers more accessible by aggregating metadata corresponding to 18 million pages and making them available via Europeana, Europe’s digital cultural heritage platform [3].
- Identify best practices for the aggregation, refinement and presentation of digitized historical newspapers.
- Develop a newspaper portal with extensive functionality for exploring digitized newspapers, for example by full text search, date, title, country of publication or language.
- Design a metadata model based on established standards like METS [4] and ALTO [5] that can serve as a best practice model for digitized newspapers.

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

III. Initial Considerations

A. Selection of suitable materials

In order to achieve the above goals of the project, a suitable subset of already scanned material had to bet selected from the different collections of the content-holding project partners. This subset should lend itself to refinement with current technology within the resource constraints (time and financial) allocated to the project.

Several challenges were immediately obvious. For instance, the material would come from different libraries (at different geographic locations and stipulating different access regulations), it would be in different languages, different layouts, and scanned at different times with different digitization parameters. It was therefore crucial to analyze all those issues and identify the corresponding technical challenges in order to create a successful refinement workflow.

In order to better understand the material available within the consortium, an initial test dataset was assembled. This initial dataset consisted of 100 pages per content holding partner, with the following selection criteria applied:

- Sample from existing (scanned) newspaper holdings.
- Representative of the quality of digitization (e.g. digitization from microfilm vs. paper originals)
- Representative in terms of file formats and file sizes
- Representative coverage of scripts and languages present in the newspaper collection

Gathering and analyzing this initial dataset proved to be extremely valuable in that it already made apparent the diversity across such a large European newspaper corpus. It also helped shape the selection criteria for the 10 million pages to be refined with OCR and OLR. Several titles from the initial dataset were subsequently deselected, mainly due to insufficient quality of the scanned images, which would likely have yielded very poor results from the refinement. It also revealed that the full dataset would feature content from as early as 1618 up to the 1990s, and with great variation regarding key aspects relevant to the OCR and OLR processes such as language and script, but also file formats:

- More than 40 languages had to be covered, including historical spelling in old newspapers and sometimes a mix of multiple languages within a single newspaper.
- Scripts that had to be tackled include mainly Antiqua (Roman), but also a large amount of Fraktur (blackletter or "Gothic") as well as Cyrillic, Hebrew and Arabic (Ottoman).
- Most common file formats for the scanned images were TIF, JPEG and PDF, but also JPEG2000 and DjVu files were present in the dataset.

This initial data sample from consortium members was complemented by a Europe-wide survey [6] that identified the wider extent of newspaper digitization in national, university and research libraries across Europe.

This was especially important as the project aimed to integrate additional newspaper content from a number of associated partners during the second half of the project, and the main requirement that had to be satisfied for this additional content was the availability of the newspaper metadata under a Creative Commons license, and wherever possible, the public domain status of the actual images and text.

B. User needs and legal implications

The results of the survey show that access to digitized newspapers is nearly always free of charge. Of the 47 institutions that responded to the survey, at least 40 (85%) offered free access to their digitized newspapers. Very few made use of the opportunity to charge users. One library had pay per view, whilst another three offered subscription services for users (i.e. paid access per day or per month).

36% (17 out of 47) of libraries have not used any form of Optical Character Recognition (OCR), meaning that searching through the full text of newspaper content is not possible. And while 64% have used OCR, only 17 of the libraries (36%) exposed the resulting full text to the viewer, indicating that they had reservations about the quality of the OCR text.

Furthermore, the survey results confirm the highly problematic access conditions for twentieth-century content. 27 out of the 47 (57%) libraries surveyed are applying a cut-off date beyond which they will not provide access to any digitized newspaper title online.

Unfortunately European law still lacks harmonization in the area of copyright, so some variation can be observed (see Fig. 1) in the dates that are applied at various organizations: 11 out of the 27 libraries with cut-off dates use a moving wall of 70 years, i.e. at the time when the survey was conducted in 2012, they would provide online access only to newspaper titles that were published in 1942, or earlier. Whenever more recent newspapers have already been digitized that are still subject to copyright, they can typically be accessed on the library premises only, or there are titles of considerable prestige and relevance, for which it was possible to arrange exceptions or agreements with rights holders on a case by case basis.

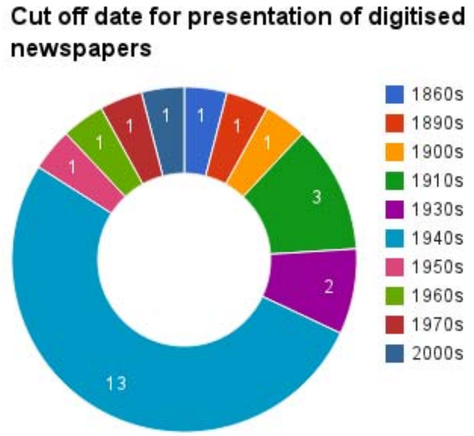


Fig. 1. Cut-off date for presentation of digitized newspapers

Accordingly, wherever possible, the selection of newspapers for refinement aimed to prioritize those titles which were in the public domain, and where the resulting full text could thus be published without any re-use restrictions.

Especially researchers aiming to explore digitized newspapers for purposes of data/text mining typically want the possibility to harvest the whole data, so that they can process it locally with specialized tools, in their own environment. For this, they need not only be enabled to download the data in the first place, but they need also be made well aware of any limitations due to licensing or copyright that could prevent them from redistributing their findings or data that they have derived from the original sources.

Furthermore, it is important to be very clear about possible biases that are introduced mainly due to the access conditions of the data. For example, an important newspaper title may be missing from the dataset due to restrictive licensing, which makes it difficult for researchers to draw valid conclusions [7].

IV. Refinement

The refinement of more than 10 million historical newspaper pages from 12 libraries required the establishment of a very strict and standardized approach.

All libraries were asked to provide information about the material they selected for refinement into a central Master List on the project extranet. This list contains descriptive or bibliographic metadata (Project-ID, Library ID such as shelf mark or similar, Title, Years of publication) but also all the technical information necessary for the refinement process, such as language, image size, font type, file type, metadata format, average file size and total number of pages.

Based on the information gathered in the Master List, the refinement workflow and a data delivery specification were defined and a workshop was held where these were discussed with all partners and agreed upon. Later on, when libraries were completing the information in the Master List, it showed that in many cases the level of granularity of the information required could not easily provided, as library catalogues typically hold no information about font types or other technical metadata that is of great importance for refinement.

A. Prerequisites

Overall there were three preparatory steps that libraries had to perform in-house before their data would be ready to be sent to the partners responsible for the refinement process.

1) File and directory structure

In order to guarantee the completion of the processing within the allocated timeframe of only 18 months, a high degree of conformity had to be guaranteed in the delivery of data to the refinement partners. A main requirement was thus the delivery of data corresponding to a strict convention for file and directory structure and names.

A specification for the delivery package structure was defined that required a root directory with a unique identifier for each newspaper title, which, in Europeana Newspapers, also encoded the ID of the holding institution. Within this directory needed to reside a subdirectory which included in its name the date of appearance of the newspaper issue in the format YYYMMDD. Not only did this help in tracking the progress of the refinement for each newspaper issue, it was also useful for generating the corresponding structure in the metadata, which was used for the (browser) calendar function. This subdirectory then held all the files, such as master and presentation images and metadata.

2) Binarisation

Based on the information gathered from the initial dataset and the Master List, it quickly became apparent that the whole dataset of images to be refined would amount to roughly 420 Terabytes of data. This revealed not only a computing but also a logistics problem – it was not conceivable to send hundreds of hard disks around Europe and keep track of them throughout the entire refinement and at the same time guarantee a smooth running refinement process.

A solution was found in that the images that were selected for refinement would be converted to bitonal (black-and-white) beforehand. Since several binarisation algorithms are available (e.g. Otsu, Sauvola, Niblack etc.) and binarisation means a loss of image information, an algorithm had to be found that had next to no negative effect on the subsequent OCR process while still delivering a significant reduction in file size.

The GPP algorithm [8] is optimized for text processing and was eventually selected as the most suitable for the given purpose. The application of the GPP binarisation to the input images reduced the file size in most cases by up to 90 percent, whilst objective evaluation confirmed the OCR quality hardly suffered from this [9] – about a 1% word accuracy reduction.

3) Image file format

The portal for presenting the digitized newspapers online required that images were provided as JPEG2000 images with tiles. This is due to the use of the open source image server IIPImage [10] which is highly suited to serving large images which require zooming. It also provides support for the International Image Interoperability Framework (IIIF) [11], a recent effort to increase the standards-based retrieval and manipulation of image data via URL.

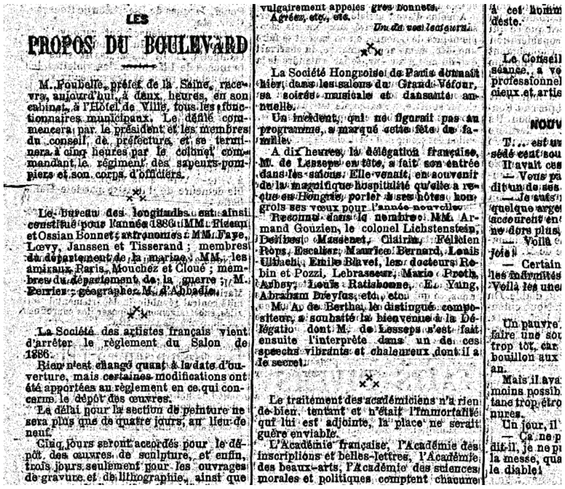
Accordingly, in a final step, libraries had to apply a tool to convert their master images to the JPEG2000 format required for presentation on the portal. Several tools and settings were tested before it was decided to use the Kakadu library [12] for which the exact parameters of the conversion were specified.

B. Processing

The processing of the newspapers was organized in two distinct workflows. The University of Innsbruck was responsible for the treatment of 8 million pages with OCR, while Content Conversion Specialists GmbH (CCS) had to process another 2 million pages with OLR.

For the 8 million pages to OCR, ABBYY FineReader Engine SDK [13] was used because it provided the greatest flexibility in configuration while maintaining aptitude for large-scale processing. The version of ABBYY FineReader used was at all times the most recent production release - at the beginning of the project v10 was available and after the first half of the project v11 was released and accordingly used.

It is in principle very difficult to forecast the average recognition time per page for such a massive refinement with several partner libraries and many different newspaper titles. Consequently, predictions were made at the start of the project, using a sample set concerning newspaper size, character number, text type and language. Already then it could be observed that the project had to deal with newspaper sources widely varying in characteristics and quality. For example, the French collection, of which the majority page format was A2, included pages that could be slightly dirty, which increased the number of “characters” on the page due to misrecognition (see Figure 2). Hence, the average recognition time with 4 servers with 8 cores each was around 10 seconds per page.



LES

PROPOS DU BOULEVARD

M. Poubelle, préfet de la Seine, recevra, aujourd'hui, à deux heures, en son cabinet, à l'Hôtel de Ville, tous les fonctionnaires municipaux. Le défilé commencera par le président et les membres du conseil de préfecture, et se terminera à cinq heures par le colonel commandant le régiment des sapeurs-pompiers et son corps d'officiers.

⁂

Le bureau des longitudes est ainsi constitué pour l'année 1886 : MM. Fizeau et Ossian Bonnet ; astronomes : MM. Faye, Lœvy, Janssen et Tisserand ; membres du département de la marine : MM. les amiraux Paris, Mouchez et Cloué ; membres du département de la guerre ; M. Perrier ; géographe : M. d'Abbadie.

⁂

La Société des artistes français vient d'arrêter le règlement du Salon de 1886.

Rien n'est changé quant à la date d'ouverture, mais certaines modifications ont été apportées au règlement en ce qui concerne le dépôt des œuvres.

Le délai pour la section de peinture ne sera plus que de quatre jours, au lieu de neuf.

Cinq jours seront accordés pour le dépôt des œuvres de sculpture, et enfin, trois jours seulement pour les ouvrages de gravure et de lithographie, ainsi que

vulgairement appelés gros bonnets.

Agréez, etc., etc.

Un de vos lecteurs.

⁂

La Société Hongroise de Paris donnait hier, dans les salons du Grand-Véfour, sa soirée musicale et dansante annuelle.

Un incident, qui ne figurait pas au programme, a marqué cette fête de famille.

A dix heures, la délégation française, M. de Lesseps en tête, a fait son entrée dans les salons. Elle venait, en souvenir de la magnifique hospitalité qu'elle a reçue en Hongrie, porter à ses hôtes hongrois ses vœux pour l'année nouvelle.

Reconnu dans le nombre : MM. Armand Gouzien, le colonel Lichtenstein, Delibes, Massenet, Clairin, Félicien Rops, Escalier, Maurice Bernard, Louis Ulbach, Emile Blavet, les docteurs Robin et Pozzi, Lebrasseur, Mario Proth, Arbey, Louis Ratisbonne, E. Yung, Abraham Dreyfus, etc., etc.

M. A. de Bertha, le distingué compositeur, a souhaité la bienvenue à la Délégatio dont M. de Lesseps s'est fait ensuite l'interprète dans un de ces speechs vibrants et chaleureux dont il a le secret.

⁂

Le traitement des académiciens n'a rien de bien tentant et n'était l'immortalité qui lui est adjointe, la place ne serait guère enviable.

L'Académie française, l'Académie des inscriptions et belles-lettres, l'Académie des beaux-arts, l'Académie des sciences morales et politiques comptent chacune

à cet homm
deste.

Le Conseil
séance, a vo
professionnel
cieux et artis

NOUV

T... est un
sédé cent sou
Il avait ces
— Vous pa
dit un de ses
— Je suis
quelque arge
accourent en
ne dors plus,
— Voilà
joie !
— Certain
les infirmités
Voilà les une

Un pauvre
faire une sou
trop tôt, car
bouillon aux
an.
Mais il ava
moins possib
tane trop étro
nures.
Un jour, il
— Ça ne p
dit-il, je ne p
la messe, qua
le diable !

Fig. 2. Example of noise in digitized French newspapers

This means that with only that kind of pages, the project would have needed over 80 million seconds for refinement, exceeding the available time frame by over 25 %. So the number of characters on a page is one important calculation factor for the planning of any digitization project, while the quality of the page is another.

However, more factors had to be considered. If one page contains Antiqua, as well as Gothic text type, ABBYY FineReader can handle both in the same recognition process automatically, but needs more time to do this. The same is valid for more than one language. Each additional language degrades the refinement speed a little bit. Having more than 4 languages at the same time slows down the process significantly. In addition, during the start-up phase of a digitization project, the decision should be made if pages get rotated automatically by the OCR engine or not. If the answer is yes than the consequence is that the OCR coordinates of the rotated page either do not match the original page coordinates or that the rotated page must afterwards be regarded as a new 'original'. Also splitting double pages produces new facts. These decisions often trace a vicious circle of actions and errors and were therefore omitted for this project, while for smaller scale projects this could be valuable.

The 2 million pages selected for OLR were processed by CCS using their in-house docWorks software technology [14]. DocWorks is an application for the conversion, structuring, and indexing of printed or electronic documents such as books, journals, newspapers and magazines. DocWorks was also used to segment newspaper pages into individual articles, which remains a highly desirable, but difficult task to achieve on complex historical newspaper layouts [15] [16].

After raw data verification and ingest, the conversion process starts off with page analysis to determine page frames, followed by several zoning and structure recognition steps, where each element is assigned a specific zone category, and the individual elements (e.g. headlines, text blocks, illustrations) are grouped together into articles.

Each automatic detection step can also be followed by manual verification, depending on the quality level required. For mass digitization projects, manual verification is typically reduced to a minimum, but the content holders participating in the OLR workflow were provided with three alternative solutions for manual quality assurance by CCS, thus allowing them to at least sample the quality, perform some manual corrections and understand the importance of good raw material for optimal automatic results.

Sufficiently good OCR results are also required for the final, experimental step that was part of the refinement workflow, named entity recognition (NER) [17]. Since NER heavily depends on language, it was decided to focus on three languages only, but which together still amount to about half of the total 10 million pages content: German, French and Dutch.

For each of the three languages, a dataset of 100 pages was selected and all entities of the type person, location and organization were manually annotated. The annotated data was then used to generate training materials for the machine learning tool used by the project, the Stanford Recognizer [18].

While the large-scale processing of the full text with NER was beyond the scope of the project, the experimental results were most promising with precision scores in the 90 percent range and recall figures somewhat lower at around 70 percent.

The significance of NER results is very considerable. An analysis of the log files of the digital newspaper archive of the National Library Wales indicated that roughly 9 out of 10 queries are for person names or locations [19]. Therefore, the availability of training corpora for NER for three languages is expected to be of great value to the NLP community and shall contribute to further enhance and extend named entity recognition systems for historical texts.

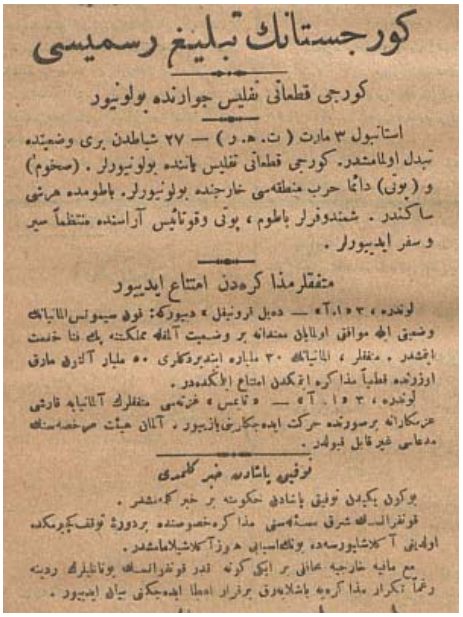
C. Challenges

Next to the sheer volume and variety of the source material, there were some particular challenges worth highlighting.

While layout analysis and article segmentation produced acceptable results, the process almost broke every time a newspaper with large tables was encountered. This is particularly problematic for newspapers that include e.g. stock market information. In the best case scenario, ABBYY FineReader simply analyses these pages for several times longer than normal pages, but in the worst case the engine stops and recognition fails. The workaround was to switch the OCR engine into a fast mode where recognition works 2.5 times faster but OCR quality decreases slightly. Other pages, not working with this setting, were recognized without table recognition – just the plain text.

Another special challenge were the newspaper holdings of the partner National Library of Turkey. The National Library of Turkey provided half of their pages with Latin text type but the other half containing Ottoman characters (resembling Arabic). Recognition for the latter one is not yet supported by the ABBYY Software and trying to process the pages with 'Arabic' produced poor output with only 20 % recognition rate.

Experiments were conducted using other OCR engines, as e.g. Tesseract version 3 features a "Cube" mode to support Arabic script. The results that were obtained from this were however not better than what could be achieved with FineReader. Accordingly, only layout recognition from FineReader was performed for the Ottoman newspaper pages, providing at least a starting ground for the future transcription or recognition of the Ottoman newspapers.



کورجستانك تبليغ رسميسی

کورجی قطعاتی تفلیس جوارنده بولونیور

استانبول ۳ مارت (ت. ه. ر) — ۲۷ شباطدن بری وضعیتده
تبدل اولمامشدر. کورجی قطعاتی تفلیس یاننده بولونیورلر. (صخوم)
و (پوتی) دائما حرب منطقه‌سی خارجنده بولونیورلر. باطومده هر شی
ساکندر. شمندوفرلر باطوم، پوتی وقوتائیس آراسنده منتظماً سیر
و سفر ایدیورلر.

متفقلرمذاکره‌دن امتناع ایدیور

لوندره، ۳ ۱۰ آ ه — دیلی قرونیقل، دیورکه: فون سیمونس المانیانك
وضعیله موافق اولمایان متمدنانه بر وضعیت آلمقله مملکته پك فنا خدمت
ایتمشدر. متفقلر، المانیانك ۳۰ ملیاره ایندیردکلری ۵۰ ملیار آلتون مارق
اوزرنده قطعیاً مذاکره ایتمکدن امتناع ایلمکده‌در.
لوندره، ۳ ۱۰ آ ه — «تایمس» غزته‌سی متفقلرك آلمانیایه قارشی
عزمکارانه بر صورتده حرکت ایده‌جکلرینی یازییور. آلمان هیئت مخصوصه‌سنك
مدعاسی غیر قابل قبولدر.

توفیق پاشادن خبر کلمدی

بوکون یکیدن توفیق پاشادن حکومته بر خبر کلمه‌مشدر.
قونفرانسك شرق مسئله‌سی مذاکره‌خصوصنده بردوره توقف کچیرمکده
اولدیغی آکلاشیلیورسه‌ده بونك اسبابی هنوز آکلاشیلامامشدر.
مع مافیه خارجیه عائلی بر ایکی کونه قدر قونفرانسك یونانلیلرك ردینه
رغماً تکرار مذاکره‌یه باشلایه‌رق برقرار اعطا ایده‌جکنی بیان ایدیور.

Fig. 3. Clipping of Turkish newspaper in Ottoman language

D. Evaluation

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 [9][20].

The bag of words evaluation per language shows that most major languages are in the region of 80% and better while there are 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).

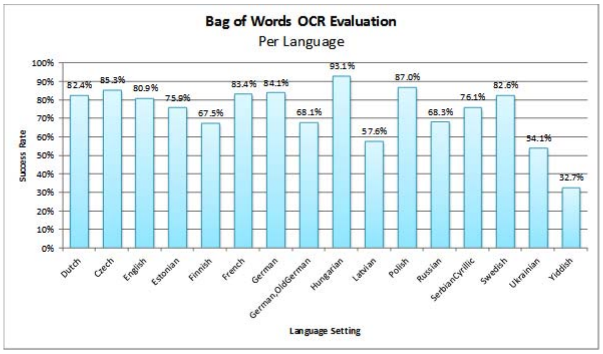


Fig. 4. Bag of Words OCR evaluation scores per language

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.

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 and reading order 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 spans two merged columns rather than following the line breaks within each individual column.

E. Recommendations

It is in principle very difficult to forecast and plan for all the possible issues that can occur in such a mass refinement project with several partner libraries and many different newspaper titles. Accordingly, the main lesson learned was that the better the preparation of the material, the better the automatic results and the lower the number of in-process changes required, which translates into fewer delays and less costs.

- Comprehensive planning and detailed project specification are vital elements of successful in-house or outsourced OCR projects. For good communication and effective project control, dedicated project managers with clear responsibilities should be in place. A collaboration platform should be foreseen to enable fast and transparent communication.
- The benefit of a highly standardized delivery data structure and its validation with checksums cannot be stressed enough. It meant that much fewer of the typical problems that normally crop up at the start of production occurred during the workflow, requiring significantly less of the expensive correction and re-processing efforts typically occurring in large projects.
- To ensure the best possible image quality, great care needs to be taken in the creation and selection of page images. If storage space and the data transfer scenario are not a limiting factor, greyscale and color images should be used as input for the workflow to create the best possible output quality.
- Deciding which metadata standards will be used to describe the digital object with technical, administrative and structural metadata is another critical factor. These standards should be XML-based and enable the metadata to be ingested into the institutional repository and presentation systems. The Europeana Newspapers project selected METS/ALTO as the open XML-based metadata standard to describe the digital object. These standards are maintained by the Library of Congress and widely used in the cultural heritage community. The highly standardized Europeana Newspapers METS/ALTO Profile (ENMAP) [21], defined during

the project, enables long-term preservation as well as data exchange and interoperability.

- Finally, deciding beforehand what types of output formats will have to be generated for the whole process is of great importance, because making intermediate changes to the formats can lead to outliers in the total collection or even require a costly re-processing of parts of the data. Especially when using a format that is newly developed or that will be developed within the lifetime of the project, it is recommended to experiment with test batches in order to finalise the format before starting the bulk processing. Here it is also extremely important to take into account what the presentation portal will look like, as decisions made for the development of such a portal, like image specifications or format requirements, will have an impact to the overall digitization workflow.

V. Concluding Remarks

Over the past three years, the partners of Europeana Newspapers project have overcome the challenge to refine over 10 million digitized historical newspapers. For the first time OCR and OLR workflows for historical newspapers were implemented on such a large scale and across a dozen European countries and many more languages. At the end of the project, almost 12 million pages have been processed and are fully searchable on The European Library [22].

A representative dataset of historical newspapers from the project - including Ground Truth - can be accessed on the PRImA website [23], while the public domain full-texts produced are continuously being released for download on Europeana Research [24]. Over the course of 2016, further data including images and XML files will be made accessible via a public API. Specialised tools and corpora for Named Entity Recognition can be obtained from GitHub [25].

Experiences gained in the Europeana Newspapers Project should be extremely valuable for future large scale refinement projects. The project final report gives a comprehensive overview of all the outcomes and achievements of the project with many pointers to detailed documents and reports [26].

Finally, this paper aimed to present the real-world challenges and experiences gained from working with OCR, OLR and NER on a very large and varied collection of historic newspapers, with the intention to serve as useful input for document analysis systems researchers and developers.

References

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

[2] Project ENUMERATE, http://www.enumerate.eu/

[3] Europeana, http://www.europeana.eu/portal/

[4] Metadata Encoding and Transmission Standard, http://www.loc.gov/standards/mets/

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

[6] Europeana Newspapers Survey Report, http://www.europeana-newspapers.eu/wp-content/uploads/2015/05/ENP-Deliverable_4.1_final.pdf

[7] A. Dunning, C. Neudecker, “Representation and Absence in Digital Resources: The Case of Europeana Newspapers”. *Presented at Digital Humanities 2014*, Lausanne, Switzerland, 7-12 July 2014, http://dharchive.org/paper/DH2014/Paper-773.xml

[8] B. Gatos, I. Pratikakis, S. J. Perantonis, “Adaptive degraded document image binarisation”, *Pattern Recognition*, 39, 3 (March 2006), pp. 317-327.

[9] S. Pletschacher, C. Clausner, A. Antonacopoulos, “Europeana Newspapers OCR Workflow Evaluation”, *Proc 2015 Workshop on Historical Document Imaging and Processing (HIP2015)*, Nancy, France, 2015, pp. 39-46.

[10] IIPImage server: http://iipimage.sourceforge.net/

[11] International Image Interoperability Framework: http://iiif.io/

[12] Kakadu JPEG2000 software development kit: http://kakadusoftware.com/

[13] ABBYY FineReader Engine: https://abbyy.technology/en:products:fre:start

[14] Content Conversion Specialists docWorks: http://content-conversion.com/#docworks-2

[15] D. Hebert, T. Palfray, S. Nicolas, P. Tranouez, T. Paquet, “Automatic article extraction in old newspapers digitized collections”, *Proc First International Conference on Digital Access to Textual Cultural Heritage (DATeCH 2014)*, ACM, New York, NY, USA, pp. 3-8.

[16] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher, “ICDAR2013 Competition on Historical Newspaper Layout Analysis – HNLA2013”, *Proc 12th Int Conf on Doc Analysis and Recognition (ICDAR2013)*, Washington DC, USA, 2013, pp. 1486-1490.

[17] C. Neudecker, L. Wilms, W.J. Faber, T. van Veen, “Large-scale refinement of digital historic newspapers with named entity recognition”, *Proc IFLA Newspapers/GENLOC Pre-Conference Satellite Meeting*, Geneva, Switzerland, 13-14 August 2014.

[18] J. R. Finkel, T. Grenager, C. Manning, “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling”, *Proc 43rd Annual Meeting of the Assoc for Comput Linguistics (ACL 2005)*, pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

[19] P. Gooding, “Exploring Usage of Digital Newspaper Archives through Web Log Analysis: A Case Study of Welsh Newspapers Online”, *Presented at Digital Humanities 2014*, Lausanne, Switzerland, 7-12 July 2014, http://dharchive.org/paper/DH2014/Paper-310.xml

[20] Europeana Newspapers Performance Evaluation Report: http://www.europeana-newspapers.eu/wp-content/uploads/2015/05/D3.5_Performance_Evaluation_Report_1.0.pdf

[21] Europeana Newspapers METS/ALTO Profile (ENMAP): http://www.europeana-newspapers.eu/wp-content/uploads/2015/05/D5.3_Final_release_ENMAP_1.0.pdf

[22] The European Library Historic Newspapers: http://www.theeuropeanlibrary.org/tel4/newspapers

[23] C. Clausner, C. Papadopoulos, S. Pletschacher, A. Antonacopoulos, “The ENP Image and Ground Truth Dataset of Historical Newspapers”, *Proc 13th Int Conf on Document Analysis and Recognition (ICDAR2015)*, Nancy, France, 2015, pp. 931-935.

[24] Europeana Research: http://research.europeana.eu/itemtype/newspapers

[25] Europeana Newspapers GitHub: https://github.com/EuropeanaNewspapers

[26] Europeana Newspapers Final Report: http://europeananewspapers.github.io/