Improving the quality of the text, a pilot project to assess and correct the OCR in a multilingual environment

Date: 31.7.2017

Author: Yves Maurer

National Library of Luxembourg, 37, bvd Konrad Adenauer, L-2450 Luxembourg

yves.maurer@bnl.etat.lu

https://github.com/ymaurer

Abstract

The user expectation from a digitized collection is that a full text search can be performed and that it will retrieve all the relevant results. The reality is, however, that the errors introduced during Optical Character Recognition (OCR) degrade the results significantly and users do not get what they expect. The National Library of Luxembourg started its digitization program in 2000 and in 2005 started performing OCR on the scanned images. The OCR was always performed by the scanning suppliers, so over the years quite a lot of different OCR programs in different versions have been used. The manual parts of the digitization chain (handling, scanning, zoning, …) are difficult, costly and mostly incompressible, so the library thought that the supplier should focus on a high quality level for these parts. OCR is an automated process and so the library believed that the text recognized by the OCR could be improved automatically since OCR software improves over the years. This is why the library has never asked the supplier for a minimum recognition rate

The author is proposing to test this assumption by first evaluating the base quality of the text extracted by the original supplier, followed by running a contemporary OCR program and finally comparing its quality to the first extraction. The corpus used is the collection of digitized newspapers from Luxembourg, published from the 18th century to the 20th century. A complicating element is that the corpus consists of three main languages, German, French and Luxembourgish, which are often present on a single newspaper page together. A preliminary step is hence added to detect the language used in a block of text so that the correct dictionaries and OCR engines can be used.

Introduction

When the National Library of Luxembourg started its text-based digitization projects in 2005, the project team knew that the quality level of the Optical Character Recognition engines was not great for historical typography. In-house trials with commercial software had resulted in low accuracy recognition, especially for gothic fonts. The library did not have a digitization lab in-house and did not have the manpower to start one, so it has always used public tenders to digitize its materials and produce the metadata.

The first larger project that the library did with text recognition was the digitization of the daily "Luxemburger Wort" in 2006. The issues that would be scanned were published between 1848 and 1950 and were overwhelmingly in German which meant that they used a gothic typeface. The OCR engine used at the time was Abbyy Finereader 8.1 and it delivered low accuracy even though it was probably the best commercial OCR engine on the market at the time.

The next larger project happened in 2007 the digitization of the daily "Tageblatt" which was scanned for the period 1913 to 1950. In both projects, the cut-off date of 1950 is related to copyright and not to the end of the publication history. In fact, both newspapers are still published today. This project was done with a different supplier who had developed an in-house OCR engine and used a voting system between 2 OCR engines (Tesseract and the in-house system) to get the best result.

In the years since, the library has worked with different digitization suppliers who have used different OCR software and the results have so far never been compared or measured. Inspired by (Reynaert, 2005), an automatic approach seemed feasible and a few good pointers from M. Reynaert got the author started.

The corpus

All of the newspapers that have been used in this paper are available are available online at www.eluxemburgensia.lu, the National library of Luxembourg's digital archives. The newspapers have been scanned and OCRed between 2006 and 2015 and the publication history is between 1841 and 2007. A total of 60791 newspaper issues, 356 000 pages, 3 million articles and 950 million words have been analyzed.

All data has been produced as METS[1] / ALTO[2] files and the pages images are in uncompressed TIFF. Starting with the first OCR project in 2005, the library has asked for article-level segmentation of the newspapers. This means that individual blocks of text are linked together to form an "Article" which can span several text columns or even several pages in the newspaper. This is achieved through a rich logical structMap inside the METS and is described in detail in the public tender documents[3].

Luxembourg has three official[4] languages and most people are at least trilingual. This means that many newspapers take a very liberal approach to language and publish articles and advertisements in different languages in the same issue and sometimes even on the same page. In Figure 1: Luxembourgish, French

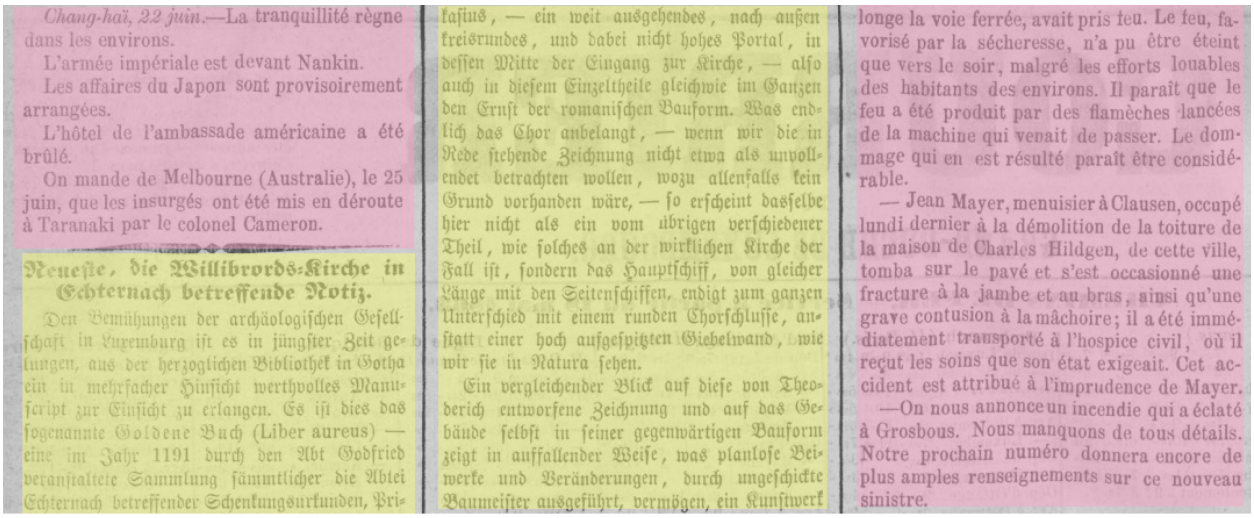
[1] http://www.loc.gov/standards/mets/
[2] http://www.loc.gov/standards/alto/
[3] http://downloads.bnl.lu/tend2016/
[4] http://www.luxembourg.public.lu/en/le-grand-duche-se-presente/langues/index.html

and English on a page there is an example of a part of a page of the Tageblatt of 23rd of June 1945 where blue is Luxembourgish, red is French and green is English. Although this example comes from the advertisement section, this mix of languages occurs throughout the newspapers and is not limited to content sent in by readers and advertisers.

This mix of languages makes the job of the OCR engine harder since not only the dictionaries, word-breaking rules and character sets change between languages, but also the fonts vary widely. For example, consider the use of fonts in Figure 2: Antiqua for French, Gothic for German. As a complicating matter articles in German were usually written in Gothic while advertisements in German used often Antiqua.

More often than not a single article is written in a single language though, so we can use the information from the article-level segmentation to find out which paragraphs belong together and are written in the same language.

Figure 1: Luxembourgish, French and English on a page (Tageblatt 23.6.1945 p.4)



Chang-haï, 22 juin.—La tranquillité règne dans les environs.

L'armée impériale est devant Nankin.

Les affaires du Japon sont provisoirement arrangées.

L'hôtel de l'ambassade américaine a été brûlé.

On mande de Melbourne (Australie), le 25 juin, que les insurgés ont été mis en déroute à Taranaki par le colonel Cameron.

Neueste, die Willibrords-Kirche in Echternach betreffende Notiz.

Den Bemühungen der archäologischen Gesellschaft in Luxemburg ist es in jüngster Zeit gelungen, aus der herzoglichen Bibliothek in Gotha ein in mehrfacher Hinsicht werthvolles Manuscript zur Einsicht zu erlangen. Es ist dies das sogenannte Goldene Buch (Liber aureus) — eine im Jahr 1191 durch den Abt Godfried veranstaltete Sammlung sämmtlicher die Abtei Echternach betreffender Schenkungsurkunden, Privilegien, — ein weit ausgehendes, nach außen kreisrundes, und dabei nicht hohes Portal, in dessen Mitte der Eingang zur Kirche, — also auch in diesem Einzeltheile gleichwie im Ganzen den Ernst der romanischen Bauform. Was endlich das Chor anbelangt, — wenn wir die in Rede stehende Zeichnung nicht etwa als unvollendet betrachten wollen, wozu allenfalls kein Grund vorhanden wäre, — so erscheint dasselbe hier nicht als ein vom übrigen verschiedener Theil, wie solches an der wirklichen Kirche der Fall ist, sondern das Hauptschiff, von gleicher Länge mit den Seitenschiffen, endigt zum ganzen Unterschied mit einem runden Chorschlusse, anstatt einer hoch aufgespitzten Giebelwand, wie wir sie in Natura sehen.

Ein vergleichender Blick auf diese von Theoderich entworfene Zeichnung und auf das Gebäude selbst in seiner gegenwärtigen Bauform zeigt in auffallender Weise, was planlose Beiwerke und Veränderungen, durch ungeschickte Baumeister ausgeführt, vermögen, ein Kunstwerk

longe la voie ferrée, avait pris feu. Le feu, favorisé par la sécheresse, n'a pu être éteint que vers le soir, malgré les efforts louables des habitants des environs. Il paraît que le feu a été produit par des flamèches lancées de la machine qui venait de passer. Le dommage qui en est résulté paraît être considérable.

— Jean Mayer, menuisier à Clausen, occupé lundi dernier à la démolition de la toiture de la maison de Charles Hildgen, de cette ville, tomba sur le pavé et s'est occasionné une fracture à la jambe et au bras, ainsi qu'une grave contusion à la mâchoire; il a été immédiatement transporté à l'hospice civil, où il reçut les soins que son état exigeait. Cet accident est attribué à l'imprudence de Mayer.

—On nous annonce un incendie qui a éclaté à Grosbous. Nous manquons de tous détails. Notre prochain numéro donnera encore de plus amples renseignements sur ce nouveau sinistre.

Figure 2: Antiqua for French, Gothic for German (Courrier Du Grand-Duché De Luxembourg 13.8.1863 p.2)

Measuring OCR quality

In any quality measurement, the idea is that you compare what data should be with what the data is actually. This means of course that you need to know what the correct output for a given problem is since that is the "should be" situation. This is referred to as the "ground truth" and it serves to evaluate the performance of algorithms which try to automatically generate the data. In the realm of OCR, this means that you have to have a corpus of scanned images together with the corresponding full text in a textual format. The difficulty in this approach is that you have to generate the ground truth by hand and this is a very time-consuming task. For the library this is unfeasible, so the author looked at ways to automatically estimate the OCR quality.

There have been numerous attempts to quantify OCR quality automatically. The EU-funded IMPACT project, which started in 2008 and wrapped up in 2011 had the aim of improving the state of the art in OCR for historical printed text. Among the 26 partner companies and institutions was also Abbyy who took advantage of the project to improve the accuracy of its OCR for gothic typefaces. In the course of the project the Polish partners from the Poznań Supercomputing and Networking Center did a comparison (Marcin Heliński, 2010) between Abbyy Finereader 10 and Tesseract 3.0.1 where they discovered that each system had its strengths and weaknesses and no clear winner could be determined.

In (Uwe Springmann, 2016), a dual approach is taken to estimate the quality of the OCR output. The mean character confidence, as reported by the OCR engine itself is taken together with the mean token lexicality. This latter quantity measures how far the recognized words are from dictionary words. The idea is that the likelihood that the OCR misrecognized a dictionary word is very low and so when the output contains a lot of words from the dictionary then the OCR quality is good.

According to (Zipf, 1949), the frequency of words in natural language text follows a power law. In essence, the most frequent word is present approximately twice as often as the second most frequent word, which is present twice as much as the third most frequent word and so on. This means that the growth of the number of unique words should slow down as the text gets larger. New words should occur with less and less frequency the more you have read.

This is used in (Camp, 2008) to explore the different shapes of the vocabulary growth curves of different corpora to see whether they can be used to say something about the underlying OCR quality. When there are a lot of misrecognized words, they are "new", so they add more vocabulary although in reality it should have been an existing word. This means that for badly OCRed text the vocabulary grows a lot faster than for a regular natural language text. Camp concluded that the Zipf curve is not a solid metric but only a crude one. However, given the ease of computation of the vocabulary growth, this crude measure is still very useful.

In this paper, we'll use a combination of these metrics since we'll use the vocabulary growth rate to double-check the OCR confidence and then compare both to the dictionary method. The latter method is the only one of the three which can reasonably be applied to checking whether one short OCRed text is better than another one because different OCR engines calculate OCR confidence in a different way and hence the values are not necessarily comparable between OCR engines.

A note on OCR Confidence

OCR confidence is a percentage, reported by the OCR engine itself, about how sure it is that the character or word it recognized is indeed what it thinks it is. This is by design a difficult problem to solve since the engine does not know what the correct text is, so it can only rely on heuristics and historical training to produce this confidence value. It is important not to confuse this with OCR accuracy, which measures the difference between the real text and what the OCR engine recognized.

(Holley, 2009) discusses the problems of relying on OCR confidence in depth and makes the point that most marketing speech for OCR programs talks about confidence and not correctness. To test how accurate the OCR confidence is, a comparison is made between the mean OCR confidence during a whole publication year with a measure derived from Zipf's law.

One of the conclusions of (Camp, 2008) was that the total length of the text has quite a big influence on the actual shape of the vocabulary growth curve. During tests on the data from the National library of Luxembourg, this was confirmed. Therefore, a variation of the vocabulary growth metric was taken which normalizes the length of the text by simply truncating it to its first million words. Since the aim is to compute a metric for a whole year of publication of a particular newspaper title, the first 1 million words from the 1st of January onwards can be considered representative of the whole year. If this is not acceptable, one can always shuffle all words and then consider the first million.

The vocabulary growth measure used here (VG) is simply the number of unique words divided by 10000, which gives us a number between 0 and 100. Observed values go from 6.8 for a French-language newspaper title published in 1868, 10.3 for a title in 1986 up to 30.3 for a different title published in 1919. 1919 is an interesting year because it has consistently the worst quality OCR results across several titles. This is probably due to the combination of the introduction of industrial paper and shortages of paper and ink which results in low quality originals.

One difficulty that is ignored here is that multilingual content will make the vocabulary grow faster than monolingual text. This is ignored because the measure is compared to OCR confidence within the same newspaper title and it is assumed that each title has a similar mix of languages throughout its publication history. Since the VG values are between 0.0001 (a single word is repeated all the time) and

100 (all words are different), it is hard to compare them directly to the OCR confidence. Therefore, they are inverted and rescaled so that they can be visually compared on a graph.

Let, for each newspaper title separately:

maxC = maximum OCR confidence value
minC = minimum OCR confidence value
maxVG = maximum VG value
minVG = minimum VG value
yVG = VG value of the current year

Then the rescaled version of the VG metric is given by:

maxC - (yVG - minVG)/(maxVG – minVG)*(maxC - minC)

Figure 3: Abbyy Finereader 8.1 Word Confidence gives an overview of how bad the correlation is between the two measures. It seems to be entirely uncorrelated. However Figure 4: Proprietary OCR Word Confidence shows an entirely different picture. Here, word confidence seems to be highly correlated to the VG measure. For other newspaper titles, the word confidence is better correlated with VG, but they have less data points since their publication history wasn't as long.

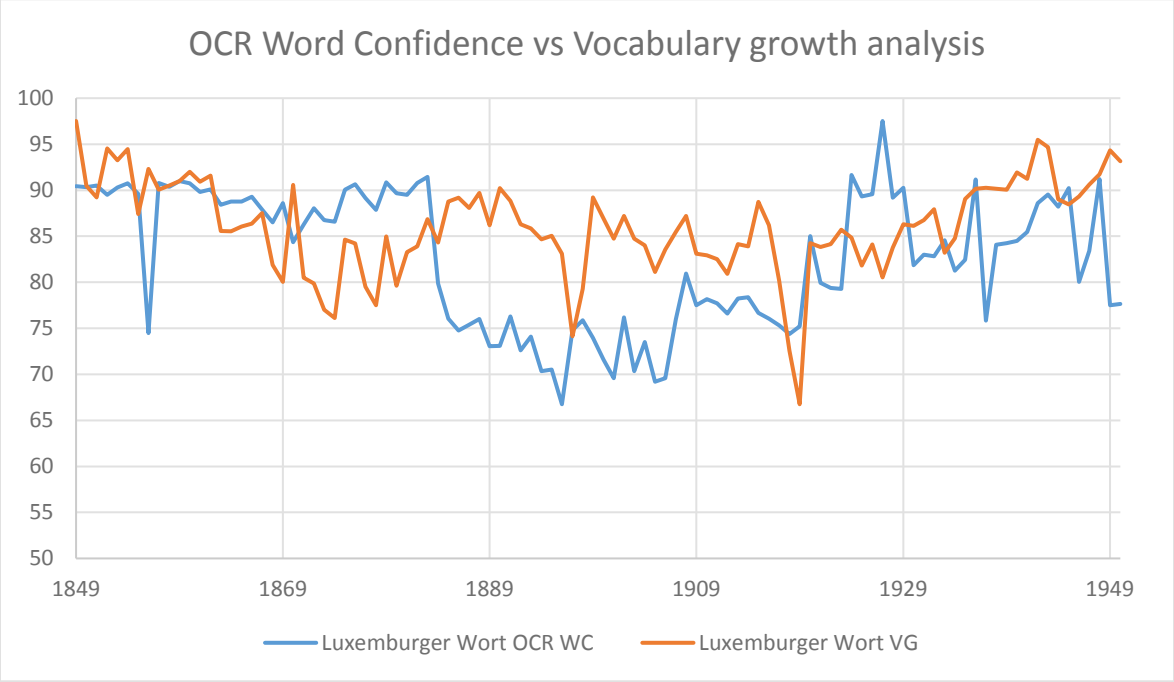


Figure 3: Abbyy Finereader 8.1 Word Confidence

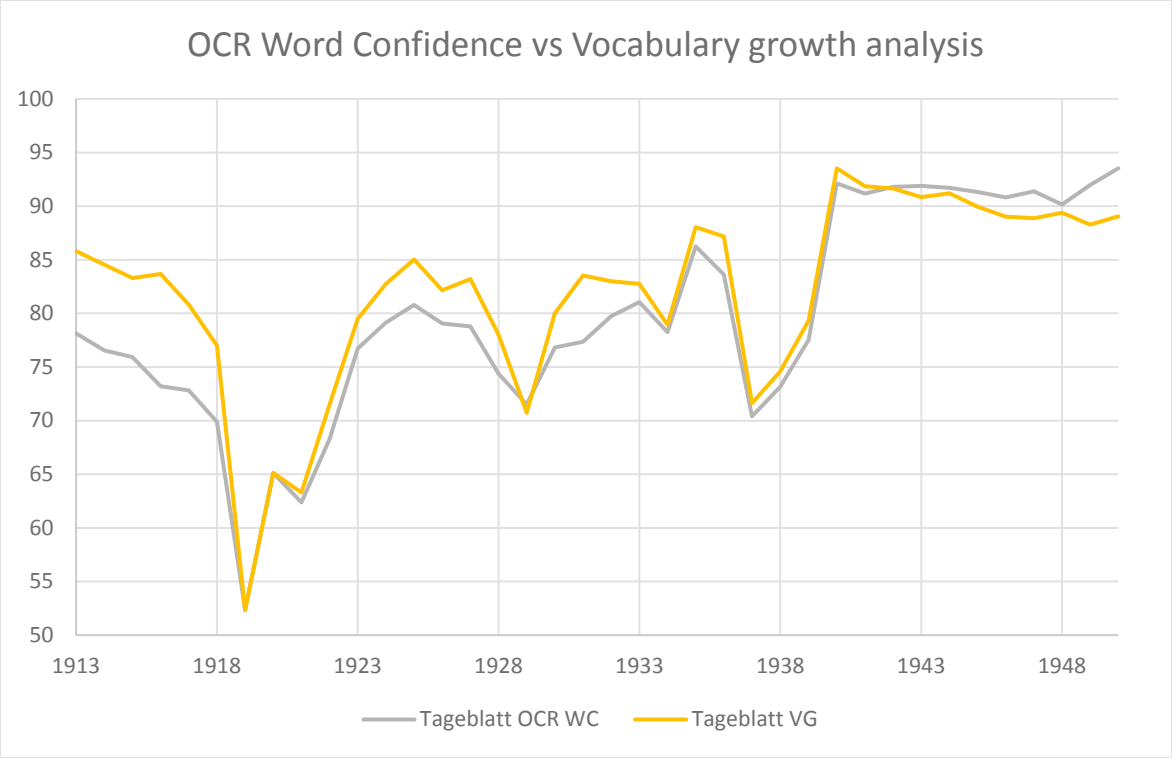


Figure 4: Proprietary OCR Word Confidence

Dictionary metric for OCR confidence

In order to calculate the VG metric one needs a million words and this is impractical when you would like to decide between the OCR outputs of two different OCR processes to gauge the best one. Hence a simpler dictionary-based metric is used. The language of the article is determined and then an appropriate dictionary for that language during the publication period is used to spellcheck the OCR result. When more dictionary words are found the OCR quality is assumed to be higher. Intuitively, this should also be correlated to the VG metric since a higher number of dictionary words probably also means that the vocabulary remains small. The decision has been taken to use the character count of the words to add more weight to longer words. This is because OCR works on a character level and hence longer words are more likely to be misrecognized and users tend to search for long words.

The Dictionary metric (DM) is calculated simply as:

nChars = number of characters in all the words in the text
nCorrectChars = number of characters in words that the spellchecker knows

DM = nCorrectChars/nChars

The spellchecker used here is hunspell[5]. It was chosen because it is a widely used spell-checking engine and has readily-available files for all the major languages in the corpus, including contemporary Luxembourgish from the spellchecker.lu project[6]. What is missing though is historical spellcheckers for

[5] http://hunspell.github.io/
[6] https://spellchecker.lu/download/libreoffice/

the different Luxembourgish spelling conventions and 19th century German. This could be solved by getting the data from the Wörterbuchnetz[7] project and compiling it into dictionaries, but this was not done in this proof-of-concept.

In order to select the correct spellchecker dictionary, the language of each article was detected using Google's CLD2[8] algorithm. It's a probabilistic naïve Bayesian classifier that uses sequences of 4 letters to detect the language. This means that its base unit is not the word but letter groups which roughly correspond to syllables and hence even unknown words can be assigned to a language. One especially striking consequence of this is that it can correctly detect the "language" of place names. The important aspects of CLD2 for this project are that it is very fast and supports all languages present in the newspaper corpus. The support for Luxembourgish is very good in fact.

In Figure 5: Luxemburger Wort WC, VG and DM, one can see that the VG and DM measures are highly correlated although different in scale. Please note that for charting purposes, DM is also rescaled so that the minimum, maximum and amplitude correspond to the same values as for the other measures.. Again, the word confidence WC from Abbyy Finereader is uncorrelated to either of the other measures.

In Figure 6: Tageblatt WC, VG and DM, the dictionary metric is correlated to both other measures which confirms that it can be used successfully to estimate OCR correctness.

[7] http://www.woerterbuchnetz.de
[8] https://github.com/CLD2Owners/cld2

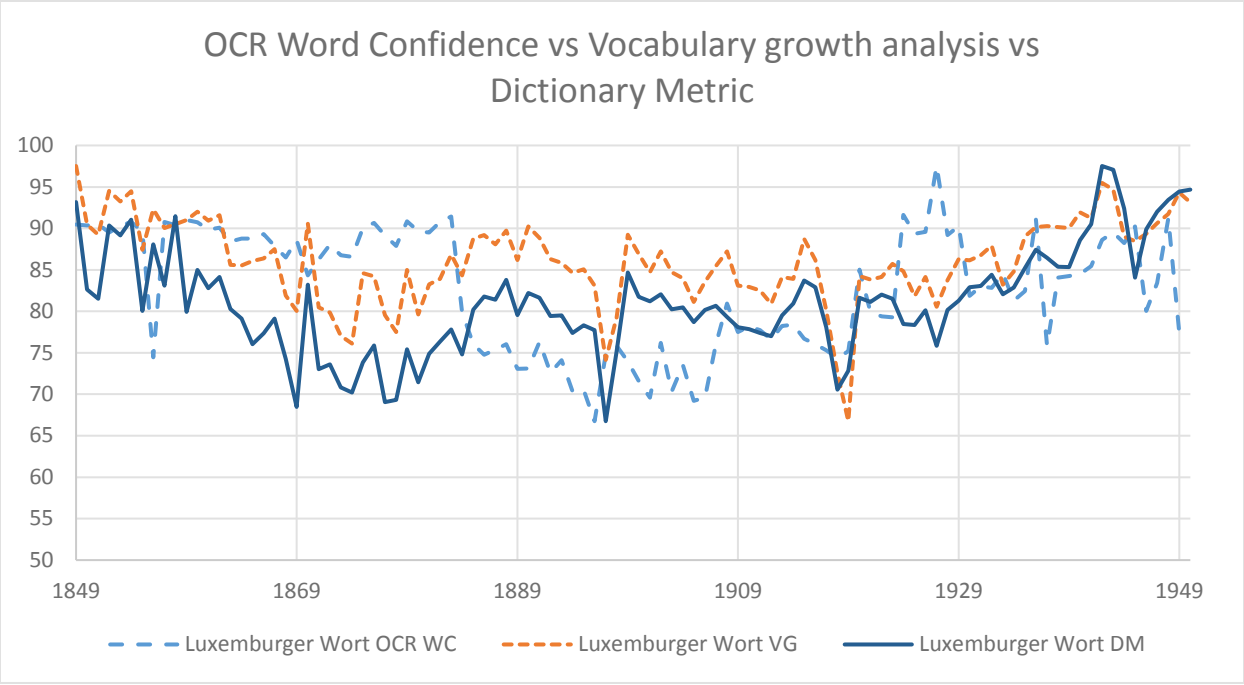


Figure 5: Luxemburger Wort WC, VG and DM

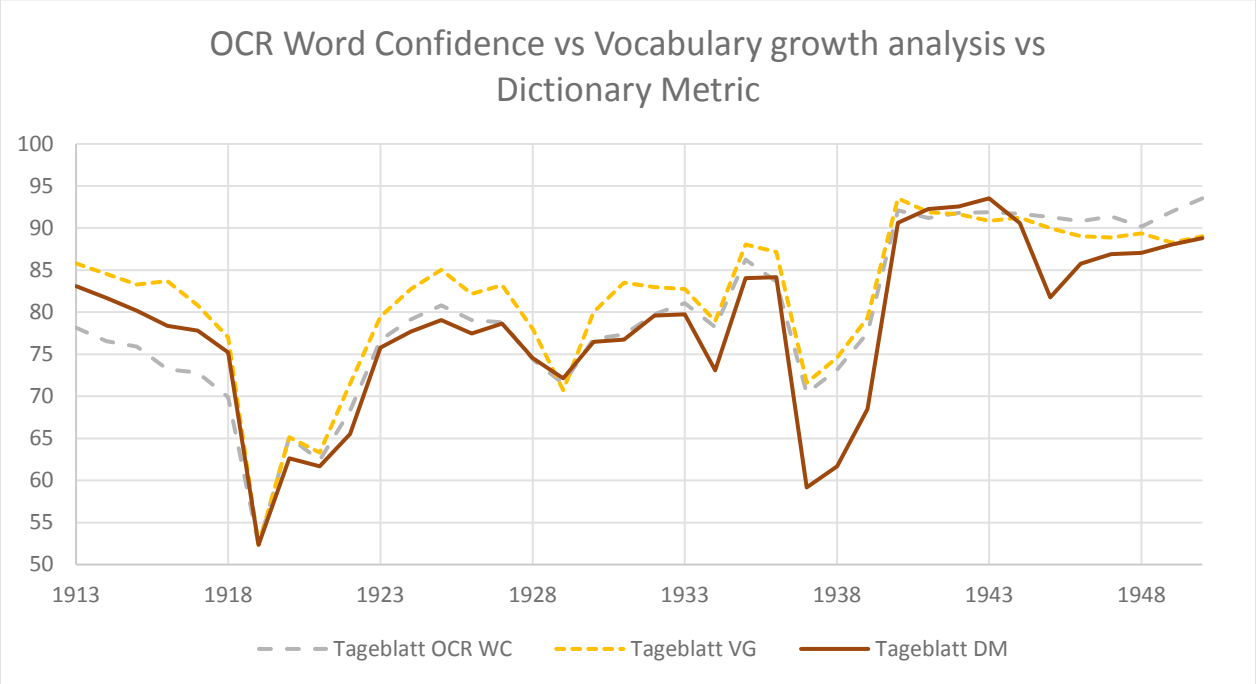


Figure 6: Tageblatt WC, VG and DM

Re-Running OCR

The following series of steps is proposed to perform a new OCR on a multilingual corpus.

1. Consider one block of text on a printed page. This is information taken from the original OCR engine or manually by the supplier.
2. Take the existing OCR text (produced by a variety of Abby, Tesseract, Omnipage, in-house OCR etc. depending on our supplier at the time)
3. Run the text through CLD2 to determine the language. This works very well for long texts, but less so for short ones.
4. If no language was detected or if it was an "exotic" language, discard the result from 3. and consider the connected blocks (i.e. the "newspaper article", "book chapter") respectively the rest of the page or document.
5. Map the language detected to the language codes used in hunspell (http://hunspell.github.io/) and try to figure out which variant to take, according to publication date, if variants are possible (e.g. German spelling variations over the centuries)
6. Segment the text into words. There could be a lot of improvement here, currently the segmentation produced by the OCR is taken.
7. Run it through hunspell and determine the dictionary metric DM1 nCorecctChars / nChars
8. Map the language to the Tesseract equivalent (and consider the period of publication for deciding whether it's blackletter or not) and run tesseract
9. Run the output again through hunspell and get the dictionary metric DM2 for the new OCR result
10. If DM2 is higher than DM1, we replace the existing OCR with the new OCR.

Since running Tesseract is a lengthy operation, only a few newspaper issues were run through the entire process. Since this is a proof of concept, care was taken to validate the approach, but less emphasis was placed upon running Tesseract efficiently. There is plenty of scope for optimizing and parallelizing the execution of Tesseract which hasn't been taken advantage of.

During experiments it emerged that step 10 is not necessarily a clear indication whether the result is better. One complicating factor is that the number of words recognized can also vary widely between the OCR engines when the image is of poor quality. In 5% of the analyzed blocks, one OCR finds at least 50% more words than the other one; roughly half of the time the original OCR finds more than Tesseract and the other times Tesseract finds more. One such example is Figure 7: poorly scanned death notice, where we get the following statistics:

Table 1: result for single death notice

	Original OCR	**New OCR**
#words	56	146
#correctWords	38	88
#characters	217	593
#correctCharacters	124	269
dictionary metric	0,57	0,45

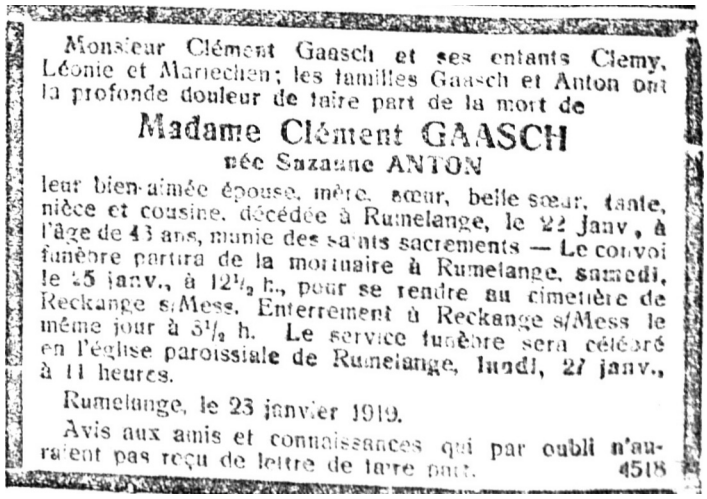
Here DM1 is higher than DM2 so the original OCR would be kept. One could argue that the fact that more words have been recognized is still more useful since then the full text search will also pick up more text. This would mean that in step 10 above our measure would be replaced by something that depends also on the number of correct words or correct characters in an absolute sense.

The raw OCR output from the original OCR is:

Mona-cur• Clément Qaasch et »e.« entant. Clemv S l Léonie et Manechen; les familles Gnasch et Anton om I ; . la ptofonde douleur dc mire part dc la mort dc | Madame Clément GAASCH « née Sazaiisie AN HON d . leur bien-aimée épouse, mère «cour, belle s«tj- <»„,<. ! Rutnelonge, le 23 jnnv.er }9|y

And the raw output from the new Tesseract run is:

Î ? . s:-'ä !“.“ l‘a“: -,m-‘s:—= Monsieur Clément Gansch et se; entame Clem . Léonie et Mancchenz les Muni îes Gsm—rh el Anton am I:) ymfunde douleur de mire part de la mont de Madame Clément GAÀSCH née Suznzme ANÏÎÛÈ‘J leur bien—uîmée épouse. mhc. sœur. beHe sœmx lanle. nièce et cousine. décédée :) Ru:nelanze, le 22 Janv, à l‘âge de 43 ans, munie des sauts sacrements — Le convoi funènre mumu de la munuaire &: Rumeîannc_ samedi. le ‘.5 ianv.. ù I2‘!, h., pour se rendre au cimenènc de Reck.mpc s;Mes‘s. Emerrement ù Reckange slMcss le même fit)… 21 W.. h. e scmcc Yunèlne sem célè.\ré nn l‘è«:hse paroxssiale de Ru=neiange, lundi, 21 ianV-à H heurcs. Rumcluug0. le ?3 janvier 10l‘J. Avis aux amis et conn m‘en! pau reçu de lo! "hi-fit} ,. _\q, \ "nî>sm.cea qui par ouh“ n'aum‘. de ln‘re 1‘m;f. 45!“



Monsieur Clément Gaasch et ses enfants Clemy, Léonie et Mariechen; les familles Gaasch et Anton ont la profonde douleur de faire part de la mort de

Madame Clément GAASCH

née Suzanne ANTON

leur bien-aimée épouse, mère, sœur, belle sœur, tante, nièce et cousine, décédée à Rumelange, le 22 Janv., à l'âge de 43 ans, munie des saints sacrements — Le convoi funèbre partira de la mortuaire à Rumelange, samedi, le 25 janv., à 12½ h., pour se rendre au cimetière de Reckange s/Mess. Enterrement à Reckange s/Mess le même jour à 3½ h. Le service funèbre sera célébré en l'église paroissiale de Rumelange, lundi, 27 janv., à 11 heures.

Rumelange, le 23 janvier 1919.

Avis aux amis et connaissances qui par oubli n'auraient pas reçu de lettre de faire part. 4518

Figure 7: poorly scanned death notice

Results

These steps have been run on 35 METS files with 194 pages from the “Tageblatt” and the “Luxemburger Wort” titles. The period convered was January 1919 and February 1937 for both cases. The reasoning was that the original OCR performed quite poorly during these periods and hence these issues would benefit most from an improved OCR result.

Tesseract did not perform consistently better than the original OCR as we had expected. In fact it performed slightly worse in an overall comparison. However, since the process is run block by block one can easily chose the best OCR result for that block (determined by the better DM score) and so obtain a combined “best-of-both-worlds” OCR. A total of 5486 blocks were analyzed and for 1993 blocks Tesseract was better or equal in terms of DM, but for 3493 blocks the original OCR had a better DM. The “Best of both” OCR gives a much better result, as can be seen in Table 2: overall results.

Table 2: overall results

	Original OCR	New OCR	Best of both
#words	642628	631372	639109
#correctWords	455206	468608	512178
#characters	3567503	3604721	3572908
#correctCharacters	2316706	2324814	2635326
dictionary metric	0,6494	0,6449	0,7376

Since the collection consists of German and French mostly, it is also interesting to check whether the OCR language has an influence on the results. As can be seen in Table 3: results for German OCR and Table 4: results of French OCR, in both scenarios the original OCR was slightly better. As expected the French OCR, which works on Antiqua fonts is better than the German OCR which works mostly with Gothic. Interestingly the DM measure is very similar for the “best-of-both” though for the two languages. This means that there was a large improvement for German while there was only a slight one for French.

In Table 5: results for German OCR on blocks for which no language could be identified, Tesseract performed much better than the original OCR. This seems to mean that it can handle noise better than the original OCR.

Table 3: results for German OCR

	Original OCR	New OCR	Best of both
#words	500406	491863	497903
#correctWords	378070	369306	407262
#characters	2822405	2869737	2832116
#correctCharacters	1996783	1884076	2155811
dictionary metric	0,7075	0,6565	0,7612

Table 4: results of French OCR

	Original OCR	New OCR	Best of both
#words	63273	62074	62816
#correctWords	51810	48877	52386
#characters	307476	302584	305633
#correctCharacters	230823	208638	234549
dictionary metric	0,7507	0,6895	0,7674

Table 5: results for German OCR on blocks for which no language could be identified

	Original OCR	New OCR	Best of both
#words	78949	77435	78390
#correctWords	25326	50425	52530
#characters	437622	432400	435159
#correctCharacters	89100	232100	244966
dictionary metric	0,2036	0,5368	0,5629

Getting the results back to the ALTO

Since we would like to improve the existing METS/ALTO files and take advantage of any manual labor that has already been done, like article segmentation, metadata enrichment, author tagging etc. it is important to keep the same text blocks and the same identifiers when doing any correction. This means that it is important to keep using the same text blocks that were used in the original OCR. In order to run Tesseract on a single block of text like this, that block is cropped from the page image into a smaller block image and this is used as input.

Then, Tesseract is set to generate an hocr[9] file, which is a standard file format, and which contains the word coordinates that are needed to generate an ALTO text block. There are several tools to perform this conversion, for example hOCR-to-ALTO[10] by Filip Kriz. The coordinates in the resulting ALTO file have to be adjusted because the OCR was run on a cropped image, and then the ALTO file can be updated.

Conclusion

The initial idea of the library that OCR correctness is not as important because the engines improve was correct, up to a point. The pipeline presented here has resulted in significant improvement of the OCR quality of the newspapers in the sample when taking the best of both OCRs. This means that if resources are made available to run it on the whole collection, the overall OCR correctness would rise and no manual labor would be required.

Some of the other factors discussed in (Holley, 2009) like original source quality and scan quality cannot be corrected for at this point. The original source material was nearly as good as possible for the "Luxemburger Wort" digitization project, but the scan quality was inferior and, moreover, the original

[9] https://en.wikipedia.org/wiki/HOCR
[10] https://github.com/filak/hOCR-to-ALTO

supplier had done “optimizations” on the images which are irreversible and degraded image quality. This means that if quality really should be improved another breakthrough in OCR technology has to happen or the library will have to re-scan the affected collection to a higher imaging quality standard.

The dictionary measure that was used to rank the OCR outputs for individual blocks has good correlation with other measures and can be easily calculated. The exact way in which OCR outputs for individual blocks are ranked in terms of quality is still not so clear though. Better measures than just taking the percentage of characters in words that are in the dictionary are needed. This is an area where more research should be done.

Bibliography

Camp, M. v. (2008). *Explorations into Unsupervised Corpus Quality Assessment.* Retrieved from ilk.uvt.nl: https://ilk.uvt.nl/downloads/pub/papers/hait/camp2008.pdf

Holley, R. (2009). *How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs.* Retrieved from dlib magazine: http://www.dlib.org/dlib/march09/holley/03holley.html

Marcin Heliński, M. K. (2010). *digitisation.eu.* Retrieved from digitisation.eu: https://www.digitisation.eu/fileadmin/Tool_Training_Materials/Abbyy/PSNC_Tesseract-FineReader-report.pdf

Reynaert, M. (2005). *Text-Induced Spelling Correction, PhD thesis.* Tilburg University.

Uwe Springmann, F. F. (2016). *Automatic quality evaluation and (semi-) automatic improvement of OCR models for historical printings.* Retrieved from arxiv: https://arxiv.org/pdf/1606.05157.pdf

Zipf, G. (1949). *Human Behaviour and the Principle of Least Effort.* Reading MA: Addison-Wesley.