On OCR ground truths and OCR post-correction gold standards, tools and formats[*]

Martin Reynaert
Tilburg center for Cognition and Communication / Centre for Language and Speech Technology
Tilburg University / Radboud University Nijmegen
Tilburg / Nijmegen, The Netherlands
reynaert@uvt.nl

ABSTRACT

We give an overview of activities undertaken in the sidelines of our automatic OCR post-correction core business over the past few years. We present ongoing projects in the Netherlands in which Text-Induced Corpus Clean-up plays a part. We describe the infrastructure we are building to help improve the overall text quality of large digitized text collections. We provide information on the tools we develop to facilitate the process and discuss the role of FoLiA XML which we adopted as a pivot format. Connecting the dots, we discuss the difference we perceive between OCR ground truths and OCR post-correction gold standards and their respective contributions.

Categories and Subject Descriptors

H. Information Systems [**H.3 INFORMATION STORAGE AND RETRIEVAL**]: H.3.7 Digital Libraries - Systems issues

Keywords

TICCL, OCR post-correction, FoLiA XML, ground truth, gold standard, evaluation

1. INTRODUCTION

In the wake of [4] we think that ad-hoc solutions and hence ad-hoc formats and scripts as a temporary fix to the problems' solutions should have had their time. In this context, we are working towards replicable and/or reproducible results based on freely available, solid, gold standards and tools to work with them and formats to store, exchange and represent them.

To this end, we here wish to describe the steps we have been and are taking to build a comprehensive platform for not only post-correction of digitized text in our main language of focus – Dutch – but also for documented and verifiable evaluations of this endeavour.

Also, in the wake of and greatly benefitting from the accomplishments in ground-truthing of OCR-texts in multiple European languages in the Impact project and in our own, much smaller, aspirations in building gold standards in two national Dutch projects we find ourself collaborating in – further described in Section 2, we aim to help develop an infrastructure capable of addressing the OCRed text problem on the vast scale it presents itself.

Most national and university libraries have at this point in time digitized sizeable selections of their paper holdings – at very considerable costs. It is the latter, in our opinion, that prompts many of these institutions to actively downplay the true extent of the problems attending the transition from paper to electronic text.

The quality of OCRed legacy texts is generally poor. Anyone who has actively looked at these, less alone those who have actually tried to use these for purposes of real digital humanities research, will not deny this fact. The figures on accuracy levels attained by the two probably most widely used OCR engines as measured within Impact, apparently measured for Polish only [5], tell their tale: there is still a long way to go.

The answer generally put forward by the people actively engaged in OCR-research and actually echoed in disclaimers such as on the Hathi Trust site[1] is that OCR technology will improve, the quality of the electronic texts will get better, all eventually will be fine.

We are not convinced, even if reOCRing probably constitutes but a relatively small part of the original cost of the digitization programmes that have been conducted. The fact is that even if OCR technology manages to attain a level of accuracy au par with the texts' ground-truths, this still not equals the digital text quality that is really required. This would be fabulous for searching, allowing for recall levels nearly as high as precision levels. This text quality would however still not be sufficient for further textual analysis or higher level linguistic enrichment or 'semanticizing' [7].

2. PROJECT CONTEXT

[*] In compliance with recent Tilburg University regulations, all the data and associated syntaxes pertinent to the current paper will be deposited at research repository `https://www.dataverse.nl/dvn/`, accessible through their DOI.

[1] Please see the help item 'Does the quality of scanned images affect the way they can be searched?' under 'Scanning/OCR Quality' at `http://www.hathitrust.org/`

DATeCH 2014, May 19-20 2014, Madrid, Spain
ACM 978-1-4503-2588-2/14/05
http://dx.doi.org/10.1145/2595188.2595216
.

In this section we sketch the current project contexts in which we develop resources and tools for OCR post-correction and its evaluation. In the course of these projects, our OCR post-correction system Text-Induced Corpus Clean-up or TICCL [9] is being completely overhauled and being largely re-implemented in C++ by our senior scientific programmer. A large scale evaluation of its performance is being prepared for a companion paper to the current one, in which we wish to focus on the language resources, corpora and tools we have been developing with the aim of making the TICCL evaluation – and hopefully others like it – possible.

2.1 NWO project Nederlab

The Nederlab project aims to build a research portal to all digitized texts relevant to the Dutch national heritage, the history of Dutch language and culture (from about A.D. 800 to the present) offering one open access, user-friendly and tool-enriched web interface, to allow scholars to simultaneously search and analyze data from texts spanning the full recorded history of the Netherlands, its language and culture. Nederlab's added value is in creating a user-friendly infrastructure for researchers, aimed at promoting cooperation and synergy, and it is hoped, at the formulation of new, often interdisciplinary, research questions.

The route followed in Nederlab is to convert all the texts incorporated into a common format, FoLiA XML [12]. In their turn all the research and analysis tools (will) have been adapted to this format. If already available online, the texts remain as they are at their original location and the linguistically or otherwise enriched versions link to these. How we manage to ensure that this is possible is explained in Section 4.3.

TICCL's role in Nederlab is to raise the quality of the OCR digitized collections to be incorporated to a higher level of quality. As a consequence, TICCL in this project is further being made diachronic, among others based on the Impact historical lexicon and named entities list for Dutch provided to us by Nederlab partner INL (Institute for Dutch Lexicology). More information about these resources is provided in [3].

Further linguistic enrichment through automatic annotation for 'lemma', Part of Speech or 'POS' and Named Entities or 'NE' are a much desired feature in Nederlab. It is a major reason why we try to improve the quality of the digitized texts we wish to incorporate in the system. Nevertheless, the post-correction will have to prove its worth, for the investment is considerable. An often overlooked because humble primary annotation step for electronic text is tokenisation, often performed in conjunction with sentence splitting, i.e. a hopefully accurate determination of where the one sentence ends and the next begins. In so far that OCRed legacy text may be noisy and/or inaccurate in its rendering of punctuation, in that it is by nature not-tokenized, we will have to address this issue in relation to OCR post-correction and the evaluation of it. This we do in Section 3.2.

2.2 CLARIN-NL project @PhilosTEI

The second project in which TICCL has a role is far more modest in scope than Nederlab. Its aim is quite simple and straightforward. Philosophers – as all other aspiring eHumanities researchers – today increasingly require high quality electronic versions of the works they study. In CLARIN-NL Call 4 project @PhilosTEI we are therefore building a work flow of web services which will allow individual researchers to upload digital images of the book's pages and receive back after processing a well formatted electronic text version fit for further building into e.g. a critical edition of the work. In the work flow, it is TICCL's task in its guise as the web service TICCLops, to enhance the text's quality, fully automatically.

This small project fits into a larger research programme[2], called 'Tarski's revolution'. The works studied in fact present a cross-section of European languages: German, French, Italian, Polish, etc. As a consequence of this project, TICCL will have been made multilingual in the sense of being able to handle texts in a range of, at least, European languages.

3. BEYOND 'MERE' OCR

3.1 The case for OCR post-correction

We have in our main paper to date on Text-Induced Corpus Clean-up or TICCL [9] analysed the distribution of typographical / type setting errors and OCR misrecognition errors in terms of their Levenshtein or edit distance [6] distribution (LD) to their canonical word forms. The take-away message is that their distribution implies that Zipf's law can be reinterpreted as: 'Accidents happen, but large ones happen rarely and small ones far more frequently'. The analysis showed that in a sample of 5,047 error types culled from the digitized Dutch Acts of Parliament or SGD[3] 89,53% of the errors lay within LD 2, 96.41% within LD 3. This is good news, because given a post-correction system such as TICCL which can exhaustively and efficiently cover the search space for all character confusions that occur in a large corpus and thereby correct the very bulk of the predominantly OCR errors that occur, we think that the case for post-correction as an alternative to eventual reOCRing with improved OCR engines is clear. To get a clear idea of TICCL's actual performance, however, we need OCR post-correction gold standards.

3.2 On ground truths versus gold standards

In the section on 'Ground Truth'[4] on the Impact Centre of Competence website we find the following description of the ground truth for text of an image: "The ground truth of an image's text content [...] is the complete and accurate record of every character and word in the image. This can be compared to the output of an OCR engine and used to assess the engine's accuracy, and how important any deviation from ground truth is in that instance."

The Impact project has provided us with a treasure trove of ground truths.

Digitisation nevertheless should go and look beyond text as displayed on the image of e.g. a printed page. Electronic text is not confined to the physical limitations of say a book's dimensions or more limiting still, the width of a typical newspaper article's columns. Digital text therefore should not be subject to or even bear undue witness of split words occasioned by the page's or column's limitations. Split words should be properly restored. So should run-on words.

[2] http://axiom.vu.nl/

[3] http://www.statengeneraaldigitaal.nl/

[4] http://www.digitisation.eu/data/browse/ground-truth/

Punctuation is a further point in question. We think an OCR post-correction system should not at this point in time be evaluated on its performance with regard to punctuation marks, we think restoration of the actual words has higher priority. A text's gold standard might well be rendered in tokenized form, punctuation marks properly split from the actual word tokens if and where appropriate. Sentence splitting may have preferably been performed.

Another major issue is that manually typeset texts may be far from perfect. The Dutch 'zetduiveltje' (E: printer's devil) used to be invoked routinely, apologetically, by printers. These inaccuracies in the text need to be corrected in any gold standard for post-correction, otherwise they may lead to an inaccurately high level of False Positives [8], unfairly penalizing the post-correction system for accurately correcting true errors.

A moot point remains in our opinion how exactly to deal with historical spelling variation, whether or not in combination with untokenized punctuation before or after the word token.

Different groups dealing with historical spelling seem to adopt different strategies: [11] transcribe into modern versions so as not to have to adapt the tools. The problem is that historical spelling variants may well exhibit large LDs to their modern form, that words were lost in the language and have no modern equivalents, etc. In Nederlab we currently opt to transcribe to the nearest historical variant. The problem then is to which historical variant to transcribe and how to measure performance given that we have e.g. 37 historical variants for the contemporary 'wenkbrauw' (E. 'eyebrow', possible literal translation: 'winking brow'). This problem may be alleviated by also performing lemmatisation to the contemporary canonical form, based on the lemmata available in the INL historical lexicon.

3.3 A (single, book size) OCR post-correction gold standard for Dutch

We have in fact so far built one such gold standard, on the basis of one of the Impact Dutch ground truths, for the book known to the Dutch National Library (Koninklijke Bibliotheek, further: KB) as DPO-035[5], one of about 10,000 Dutch in the collection 'Early Dutch Books Online' or EDBO. In CLARIN-NL Call 1 project TICCLops we created a generic solution for turning linguistic applications in web services / applications called CLAM[6] (Computational Linguistics Application Mediator) and put TICCL as 'online processing system' TICCLops[7] online. The system is meant as a demonstrator around the digitized book. In order to give the user an idea of what OCR post-correction can do – and what TICCL at the time could achieve in terms of improving the book's accuracy, we manually converted the ground truth or GT into a gold standard. This became a dual gold standard: one for the actual historical text in its printed spelling – the historical gold standard or HGS, one for the contemporary version – the contemporary gold standard or CGS, although only on the level of the spelling of the individual words. This work also made apparent to us the actual difficulties involved in transcription work of this kind.

Our gold standard has 55.812 lines of lined-up OCR strings, ground truth strings, historical gold standard and modern gold standard strings in tab-separated columns. On the basis of the totaled diverging strings as shown in Tables 1, 2, 3 and 4 we determine the word accuracies. The observed word accuracy based on the gold standard of the OCR version in terms of word tokens compared to the GT is 87.54%, to the HGS it is 88.74% and to the CGS 79.85%. We contrast this to the accuracy of the HGS compared to the CGS: 85.34%. This gives us a good idea of the work to be done by e.g. a post-correction system. The higher accuracy of the HGS compared to the GT is explained by the fact that in the OCR version hyphenated end-of-line word splits are resolved. We assume that this is an active automatic post-processing step performed by the KB OCR service providers rather than by the OCR engine. We see these results as a major incentive to also pursue OCR post-correction, rather than solely trying to improve the OCR process.

The word accuracies just stated diverge from the ones we have in [10]. We there state that the accuracy of the OCR version in terms of word tokens compared to the HGS is 88.24%, a difference of 0.2%, and to the CGS 76.24%, a highly annoying difference of 3.61%. We obtained the current results on the basis of the gold standard alone, the other published ones on the basis of the script that measures our system's performance scores. The discrepancy is no doubt due to subtle differences in handling punctuation and whether or not strings that differ in punctuation only or consist of punctuation only are seen as incorrect and therefore target or not. In fact, leading and trailing punctuation marks to the word strings were removed in the current count. The OCR version suffers badly from spurious punctuation marks. We are determined to address this issue more rigorously in further work.

The main question to us is: what can be achieved by OCR post-correction? We have also gathered statistics on the distribution of the shifts in terms of LD. In the appendix we detail the full statistics about the actual shifts observed between the OCRed version of the book, its Impact GT and both the HGS and CGS we derived from them. All classes in the Tables are described from the point of view of the correct version, be it ground truth, historical gold standard or contemporary gold standards. The class 'multisingle' might as well have been called 1 to 2, 1 to 3, etc. substitutions. A typical example of these is the historical spelling variation 'g' to 'ch' in Dutch, which in contemporary Dutch resolves the ambiguity for the historical word form 'ligt' into both 'hij ligt' (E.: 'he lays') and 'het licht' (E.: 'the light'). The class 'multiple' are shifts that can only be described in terms of combinations of deletions, insertions and substitutions. The class 'space insertions' refers to split words in the OCR version. However, this does not give an exact count of the split words present, our classifier may see these as part of multiple shifts. The actual amount of split words in the OCR version is quite elevated: the gold standard lists 1,568 compared to the HGS, which also shows 195 run-on words.

Spelling correction systems, of whatever kind, have a certain 'reach' in terms of edit distance within which they operate. The sums of the totaled percentages in Table 2 shows that compared to the HGS, 88.98% of the shifts lay within LD 2, 94.91% within LD 3, for the CGS Table 3 shows 83.42% within LD 2, 93.26% within LD 3. Given sufficiently powerful and comprehensive OCR post-correction systems, most

[5] `http://resolver.kb.nl/resolve?urn=dpo:35:mpeg21:0009`

[6] `http://proycon.github.io/clam/`

[7] `https://portal.clarin.inl.nl/ticclops` provides access to the system

errors should therefore be resolvable.

4. RESOURCES

In this section we present a range of activities in the margins of our actual work on corpus clean-up we have been engaged in over the past few years.

4.1 Pivot format: FoLiA XML

We have adopted the Format for Linguistic Annotation or FoLiA XML as the pivot format for our OCR post-correction system TICCL. Working in close collaboration with the format's main developer, we have been able to extend it into a format in which we can express what is required to be able to fully describe the text of e.g. a digitized book, its OCR ground-truth and its OCR post-correction gold standard – if these are indeed available. Further it has the necessary provisions for the corrections a system such as TICCL may want to effect. Corrections may be expressed in gradations: they may be straight-on corrections, accompanied by a particular confidence score and/or a record of whether the correction was made by an automated system or manually, or they may be suggestions for corrections – an enumeration of correction candidates, ranked according to the confidence scores assigned to them.

The way this is currently implemented in the actual 'FoLiA correction module' of TICCL, a C++ program developed by our senior scientific programmer, is that up to the specified number of correction candidates or CCs are added to the suspect OCR string and that the best-ranked CC actually substitutes or 'corrects' the suspect string in a new paragraph element. For subsequent processing one may then opt to linguistically enrich the corrected paragraph identified by its XML attribute 'Ticcl' rather than the original OCR paragraph with attribute 'OCR'. For indexing, one may then very well opt to index both versions of the paragraph, preserving the original noisy paragraph as reference but enhancing search recall by the added CCs.

4.2 VU-DNC corpus as post-correction gold standard showcase

The VU-DNC[8] corpus was developed in CLARIN-NL Call 2. The corpus consists of two sets of newspaper articles drawn from 5 national Dutch Newspapers. It is diachronic, the first set comprises articles from the years 1950-1951; the second set articles from 2002. It is annotated for subjectivity and quotations, the topic of the PhD work of its compiler [13]. The 2002 set was born-electronic. The first set was digitized by means of ABBYY Finereader version 9 and was initially post-corrected in part by means of the Microsoft Office Spelling Tools.

A benchmark for OCR post-correction has been built on a collection of texts in the older Dutch spelling De Vries - Te Winkel. The pre- and post-correction versions of a larger part of the corpus have been aligned semi-automatically, by employing student-assistants using existing tools and algorithms developed at ILK. Alignment has been made at word level. The resulting annotated corpus and post-correction benchmark are available through the Dutch HLT Agency TST-Centrale [9].

4.3 To and from FoLiA XML

To the best of our knowledge there are three distinct file formats dedicated to linking OCRed text strings to their respective positions on the page image. These are Alto XML, hOCR HTML and Page XML.

Alto XML is the format used by the large digitization projects conducted by the Dutch National Library or KB. It is typically used to represent the text of a single book page or a single newspaper article and holds the positional references for each word string with respect to the image the strings were derived from.

In so far as newspapers typically have several articles distributed over a single and occasionally several newspaper pages, Alto files for single newspaper articles are referenced by another file in the DIDL-format. One DIDL then describes the entire layout of a single printed newspaper. A single, structurally simpler DIDL file may describe a book, for instance, and hold all the references to the Alto-files containing each separate page's text.

We have built the C++ program FoLiA-alto which on the basis of a single newspaper's or book's DIDL file harvests the attendant Alto XML files from the KB's repositories. After downloading, FoLiA-alto then converts the Alto XML into FoLiA XML, ready for post-correction by TICCL. The FoLiA version retains sufficient referential information with respect to the original scanned page image for preserving the query string image highlighting function provided by the Alto XML.

The HTML format hOCR [2] was developed as an open standard for OCR results and seems linked most closely to the open-source OCRopus project[10]. It is also an optional output format for the open-source OCR-engine Tesseract[11], originally developed by HP and now being further developed under the aegis of the Google Books project. We have adopted it in the CLARIN-NL project @PhilosTEI as a first and intermediate step towards a final TEI XML P5[12] formatted digitized book format fit for further – largely manual – processing into 'critical edition' of the particular philosopher's original work. In so far that @PhilosTEI also relies on an intermediate OCR post-correction step, we have built a convertor from hOCR HTML to FoLiA XML. This too is to be post-corrected fully automatically by TICCL, after which the final conversion to a basic TEI XML P5 format will be effected.

Page XML was developed in the European project Impact to fill the gaps left by both the Alto XML and hOCR HTML formats with regard to affording an exhaustive description of the printed page with respect to the various levels and dimensions of the OCR-process, be they to do with the physical conditions of the page, with the peculiar aspects of the print and the fonts involved or with the text, i.e. the groundtruthed version of the actual text as printed. With regard to the latter – and of most relevance to our concerns here – we have seen no elaborate use made of the provisions available in the Page XML specifications. The ground-truths developed in the Impact project we have been able to inspect this far all have the texts rendered in blocks, as they appear on the printed pages. There is no further, deeper, positioning

[8] http://www.clarin.nl/node/440

[9] At: https://portal.clarin.inl.nl/vu-dnc/ VU-DNC data are available to CLARIN members, the CMDI metadata to all.

[10] http://code.google.com/p/ocropus/

[11] http://code.google.com/p/tesseract-ocr/

[12] http://www.tei-c.org/Guidelines/P5/

information to the individual text strings, as is provided e.g. by the KB Alto XML files. hOCR keeps a middle ground in this, referencing full text lines, instead of the Alto XML word strings or the Page text blocks. We have built a third convertor in C++ that renders the Page XML text blocks in FoLiA XML as paragraph elements. Each string within the OCR text block, as defined by intermediate white space, then becomes a FoLiA string element, retaining the available positional information which refers back to the original page image.

4.4 The problem of aligning 'old' with 'gold'

The Impact project has produced a wealth of OCR ground truth versions of texts spanning several centuries, covering a nice range of European languages. For our purposes, however, it is a set-back that preservation of the noisy version of the OCRed texts seems not have been part of the agenda. If these were, we would now have an excellent basis to build post-correction gold standards on, preserving the original text segmentation. The point is not that we lack OCR-versions of texts, far from it. The point is that we lack OCRed text lined up with its ground truth version because it is this combination that most easily allows us to derive a proper post-correction gold standard from it and to perform the kinds of measurements we have done on DPO35.

While we now have, as explained in the previous subsection, three convertors that deliver OCR-output and ground truth versions of the particular texts in a uniform XML format geared at OCR post-correction, the essential problem of aligning an OCR version to its ground truth remains. We saw that the 2 main OCR engines and the ground truths differ, not necessarily in their segmentation of the texts, but certainly in their labeling of same. This results in different identifiers for the particular segmentation text blocks.

In the VU-DNC project we have built 'Goldie-Oldie', a word token level text aligner for gold standards and their 'old' versions, i.e. those produced by an OCR-engine. Goldie-Oldie currently is still not a FoLiA tool. It works on column formatted text. It further uses robust matching based on anagram hashing [9]. Anagram hashing provides a different, in fact numerical, representation for text strings and this bypasses regular text matching pitfalls (wildcards, reserved characters and the like), thereby further providing an elegant way of disregarding minor differences between two strings, the gold version versus its old one, such as misrecognized punctuation or punctuation noise (i.e. extra punctuation added by the particular OCR engine).

We have recently learned about a similar tool, RETAS [14] . This starts its alignment based on the neat idea that texts through their Zipf distribution typically have about 50% of their word types being hapaxes. These singletons are identified first – we take it in the ground truth version – and are then used as 'anchors', allowing for subdividing the alignment problem into the text intermediate to a pair of these anchors. Much as we like this idea, Goldie-Oldie does likewise, but starts off from all exact matches between the two text versions: if two strings in the approximate same area of the text share the same anagram value, they are regarded as 'anchor candidates' and the system will try to align the intermediate text. We have so far not been able to formally evaluate Goldie-Oldie nor to compare its performance to RETAS. This we reserve for future work. What we do observe in the aligned ground truths provided by the project, is that their definition of OCR ground truth is probably far closer to our own idea of a gold standard. The RETAS ground truths do preserve capitalization, but seem to have discarded nearly all punctuation, seem not to preserve original split words, etc.

5. NOTES ON THE EVALUATION OF OCR POST-CORRECTION SYSTEMS

5.1 Inadequacy of the current gold standard

An important aspect of OCR post-correction gold standards for diachronical texts that has become abundantly clear in the evaluations we performed for [10] is their inherent limitations with regard to the wide variety in attested historical spellings for many Dutch word forms. The bulk of the False Negatives incurred by the system tested were due to the fact that it proposed other attested historical spelling variants than the ones in the Gold Standard. A solution to this problem might be to involve the attested modern lemmata for these word forms in the evaluation as we have these available in the INL historical lexicon. This would certainly be fairer to the system with regard to its performance in light of more accurate text for retrieval purposes. But then this might not be acceptable to those who wish for the most accurate text for the purposes of further linguistic enrichment or geared to the actual study of the phenomena in the historical texts.

5.2 Reporting an upper-bound

We missed a good chance in our reporting of the evaluations in [10] and that is to state the upper boundary of what the system could achieve in light of its inherent limitations and the limits it was set to work in. Limitations are that it cannot currently deal satisfactorily with split and run-on words. A limitation of probably all post-correction systems would be that they cannot possibly deal adequately with OCR errors in numbers. Its limits were that its reach in terms of LD was 2 characters and that it was set to work on word strings from 6 characters in length to 36 characters in length. Given the available gold standard, the sum of the items surely not resolved, i.e. not possibly adequately corrected by the system, subtracted from the total of target items as defined by the gold standard gives the upper-bound attainable. We should recommend as 'best practice' stating this upper-bound in evaluation reports on text correction or normalization. Following this best-practice recommendation liberates researchers from the need (or desire, or urge) to more favourably present their systems' performance results by measuring only what it is their system is designed to help to solve. An example of this latter strategy is presented by [1]. Comparing one's results against the ceiling represented by the upper-bound frees researchers from the need to re-annotate their data in light of only the issues tackled by their system in order to get a better view of its performance, as these authors found they had to do.

5.3 Pitfalls in evaluating OCR post-correction

We have quite recently learned that evaluating OCR post-correction systems may have unexpected pitfalls. In calculating the results for [10] we were faced with rather major discrepancies between the actual counts and the counts expected on the basis of the gold standard. The non-target score showed a discrepancy very in line with the number of split words in the OCR. About 1,000 more True Negatives were counted than designated non-target. It was only after major searching for an acceptable explanation that we realised that this was indeed due to parts of split words in the OCR. The fact is that in Dutch at least, split word parts may well be non-words, but may very well represent real-words also such as the most frequent Dutch word 'de', the definite article 'the'. If in the gold standard, these may then erroneously be counted as True Negatives. For the target we had another discrepancy of about 200 items that were unaccountably missing. In the end we had to conclude that these were single non-ordinary punctuation marks that the system rightfully removed.

6. WORKING TO SCALE

As stated on the home page of the current beta version of the new KB portal[13] to its online collections the site gives free full-text access to more than 90.000 book publications dating from the 18th and 19th centuries, to 1 million newspapers from the 17th, 18th, 19th en 20th centuries, to 1.5 million pages of Dutch periodicals and magazines of the 19th and 20th centuries, and finally, to 1.5 million digitized radio bulletin type scripts covering the years 1937 to 1984. Another KB site[14] gives access to 2.5 million digitized pages of Dutch Acts of Parliament ranging from 1814 to 1995 on which we have worked in NWO project Political Mashup. As we have stated before, if we want to stand any chance of improving the overall quality of these amounts of digitized textual materials, we need the help of fully automatic OCR post-correction systems.

7. CONCLUSIONS

We have given a survey of work we have undertaken over the past years to support our work on automatic, unsupervised OCR post-correction of very large digitized text collections. Text-Induced Corpus Cleanup-up or TICCL [15] has recently been evaluated on the whole Dutch contents of EDBO and the results are to be presented shortly [10].

However tedious and time-consuming the task may be, if we are ever to fully gain the necessary insight into what a proposed OCR post-correction can actually achieve, we will need to build the required gold standards to properly evaluate the tasks. As we have explained, OCR ground truths are a good starting point to this end, provided the raw OCR versions from which they have been built and to which they are properly aligned are preserved and made available. If only this message is conveyed by the current paper to further large scale projects geared at improving digitisation, our mission here will have been accomplished.

8. ACKNOWLEDGMENTS

Martin Reynaert acknowledges support from NWO in project Nederlab and from CLARIN-NL in projects @PhilosTEI (CLARIN-NL-12-006) and OpenSoNaR (CLARIN-NL-12-013).

9. REFERENCES

[1] B. Alex, C. Grover, E. Klein, and R. Tobin. Digitised historical text: Does it have to be mediOCRe? In *Proceedings of KONVENS 2012 (LThist 2012 workshop)*, pages 401–409, 2012.

[2] T. Breuel. The hOCR microformat for OCR workflow and results. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition*, volume 2, pages 1063–1067. IEEE Computer Society, 2007.

[3] J. de Does and K. Depuydt. Lexicon-supported OCR of eighteenth century Dutch books: a case study. In R. Zanibbi and B. Coüasnon, editors, *DRR*, volume 8658 of *SPIE Proceedings*. SPIE, 2013.

[4] A. Fokkens, M. van Erp, M. Postma, T. Pedersen, P. Vossen, and N. Freire. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[5] M. Helinski, M. Kmieciak, and T. Parkola. Report on the comparison of Tesseract and ABBYY Finereader OCR engines, 2012.

[6] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8):707–710, 1966. Original in *Doklady Akademii Nauk SSSR* 163(4): 845–848 (1965).

[7] D. Odijk, E. Meij, and M. de Rijke. Feeding the second screen: Semantic linking based on subtitles. In *Open research Areas in Information Retrieval (OAIR 2013)*, Lisbon, Portugal, 05/2013 2013.

[8] M. Reynaert. All, and only, the errors: more complete and consistent spelling and OCR-error correction evaluation. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008. ELRA.

[9] M. Reynaert. Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal on Document Analysis and Recognition*, 14:173–187, 2010. 10.1007/s10032-010-0133-5.

[10] M. Reynaert. Synergy of Nederlab and @PhilosTEI: diachronic and multilingual Text-Induced Corpus Clean-up. In *Proceedings of LREC'14*, Reykjavik, Iceland, 2014. ELRA.

[11] M. Reynaert, I. Hendrickx, and R. Marquilhas. Historical spelling normalization. a comparison of two statistical methods: TICCL and VARD2. In *Proceedings of ACRH-2*, pages 87–98. Lisbon: Colibri, 2012.

[12] M. van Gompel and M. Reynaert. FoLiA: A practical XML Format for Linguistic Annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3, 2013.

[13] K. Vis. *Subjectivity in News Discourse: a Corpus Linguistic Analysis of Informalization.* Vrije Universiteit, 2011.

[14] I. Z. Yalniz and R. Manmatha. A fast alignment scheme for automatic OCR evaluation of books. In *Proceedings of ICDAR '11*, pages 754–758, Washington, DC, USA, 2011. IEEE Computer Society.

[13] http://www.delpher.nl

[14] http://www.statengeneraaldigitaal.nl/

[15] TICCL in its new C++ implementation is to be available via http://ticclops.uvt.nl/

APPENDIX

We here present the statistics obtained from the comparisons between the OCR version of DPO35, its Impact ground truth GT and both the historical gold standard HGS and contemporary gold standard CGS. We further present the same statistics as measured on the historical gold standard HGS compared to the modern, contemporary gold standard CGS. Results are presented per Levenshtein distance observed. Capitalisation and leading or trailing punctuation were not taken into account. For exact counts of split words and run-ons, please refer to Section 3.3. We there also discuss the classes in more detail.

Table 1: OCR version versus GT

Category	LD 1	LD 2	LD 3	LD 4	LD 5	LD 6 - 12	Total	%
deletion	314	906	3	4			1227	15.444
insertion	273	19	6	5		1	304	3.826
substitution	3197	375	40	5	3	1	3621	45.576
transposition		2					2	0.025
multisingle		532	212	54	41	28	867	10.913
multiple		413	419	162	57	76	1127	14.185
space deletion	72						72	0.906
space insertion	718						718	9.037
TOTAL	4578	2247	680	233	101	106	7945	
%	57.621	28.282	8.559	2.933	1.271	1.334		100.0

Table 2: OCR version versus HGS

Category	LD 1	LD 2	LD 3	LD 4	LD 5	LD 6 -12	Total	%
deletion	257	23	2	1			283	3.995
insertion	277	97	7	9	6		396	5.590
substitution	3305	387	42	2			3736	52.739
transposition		3					3	0.042
multisingle		561	123	33	37	39	793	11.194
multiple		447	247	115	45	73	927	13.086
space deletion	75						75	1.059
space insertion	864						864	12.196
TOTAL	4785	1518	421	160	88	114	7084	
%	67.547	21.429	5.943	2.259	1.242	1.581		100.0

Table 3: OCR version versus CGS

Category	LD 1	LD 2	LD 3	LD 4	LD 5	LD 6 -12	Total	%
deletion	315	22	3	1			341	2.420
insertion	3378	190	11	8	4		3591	25.490
substitution	2342	341	92	42	2		2819	20.010
transposition		7					7	0.050
multisingle		3605	514	93	59	47	4318	30.650
multiple		853	767	391	134	168	2313	16.418
space deletion	52						52	0.369
space insertion	643						643	4.564
TOTAL	6734	5018	1387	535	199	215	14088	
%	47.800	35.619	9.845	3.798	1.413	1.505		100.0

Table 4: HGS versus CGS

Category	LD 1	LD 2	LD 3	LD 4	LD 5	LD 6 -12	Total	%
deletion	148	12	1				161	1.680
insertion	3834	537	1	1			4373	45.623
substitution	438	126	15	8	2		589	6.145
transposition		2					2	0.021
multisingle		3844	16	20	9	2	3891	40.595
multiple		101	246	148	49	21	565	5.895
space deletion	2						2	0.021
space insertion	2						2	0.021
TOTAL	4424	4622	279	177	60	23	9585	
%	46.155	48.221	2.911	1.847	0.626	0.239		100.0