OCR Post-Correction Evaluation of Early Dutch Books Online – Revisited

Martin Reynaert
TiCC, CLST
Tilburg University, Radboud University Nijmegen
reynaert@uvt.nl

Abstract

We present further work on evaluation of the fully automatic post-correction of Early Dutch Books Online, a collection of 10,333 18th century books. In prior work we evaluated the new implementation of Text-Induced Corpus Clean-up (TICCL) on the basis of a single book Gold Standard derived from this collection. In the current paper we revisit the same collection on the basis of a sizeable 1020 item random sample of OCR post-corrected strings from the full collection. Both evaluations have their own stories to tell and lessons to teach.

Keywords: OCR Post-Correction, Evaluation, Digitized Books, Dutch, Diachronic

1. Introduction

In Reynaert (2014b) we announced the pending availability of a fully-automatic and unsupervised multilingual OCR post-correction system fit for diachronic work[1]. We also presented this reimplemented Text-Induced Corpus Clean-up system or TICCL applied to the post-correction of a collection of over 10,000, mainly late 18th century, diachronic Dutch books as well as an extended evaluation based on the Gold Standard for a single book from 1789 (known as DPO35).

While we stand fully by this evaluation, it can be expected that the text of a single volume – a history book specifically directed at children – does not cover the full gamut of phenomena that may occur in the larger collection. We therefore now strive to find a way of charting these phenomena and how our system manages to deal with them on the basis of a totally different, but necessarily still limited, sample. What one needs to come to grips with in work as this, is the sheer size of the search space we are dealing with. A list of over 800 million pairs of word variants paired to possible Correction Candidates or CCs is beyond any human's comprehension. We are quite incapable of even beginning to evaluate this full list. All we can possibly do is query it for a limited subselection and try to comprehend what this limited sample may tell us. This is exactly what we try to do in this paper.

In Section 2. we present a new evaluation set randomly sampled from TICCL's output on the 10K Early Dutch Books Online corpus and briefly discuss how we verified and annotated it. We next in Section 3. derive evaluation scores and further sample statistics and analyse them in some depth. The discussion in Section 4. deals with costs and benefits of the two evaluation methods, discusses specific OCR phenomena and finally offers an informed and prioritized list of possible extensions to our own – and possibly other – OCR post-correction systems.

[1]Available in the CLARIN infrastructure at `http://ticclops.clarin.inl.nl` and with a new interface at `http://philostei.clarin.inl.nl`

2. A new Gold Standard

2.1. TICCL briefly recapitulated

In order to comprehend the following it is doubtless necessary to be aware that TICCL[2] derives what amounts to a language model from the vocabulary of the lexicon and the full corpus it is set to correct as well as an error model tailored to the particular corpus. The error model is induced from the pairs of focus variants and CCs it retrieves and retains if they fall within the Levenshtein Distance (LD) limit that was set. Both language and error model have their role to play in the ranking of CCs, next to a range of lesser features.

For identifying spelling variants, TICCL relies on anagram hashing which we fully described in (Reynaert, 2010) and gave new implementation details for in (Reynaert, 2014b). The essence is that all the characters in the alphabet are assigned a large numerical value which puts them apart in Euclidean space at fixed and known distances. These distances remain as identical for two individual characters, e.g. 'c' and 'r' as for words differing only in these two characters, e.g. 'cat' and 'rat'. This extends to all possible combinations of characters given the alphabet. So, given the numerical value for any character(combination) substitution, all pairs of words displaying the same numerical value difference between the sums of the numerical values for their individual bags-of-characters, necessarily display the same character substitution(s). This allows for efficient exhaustive identification of all the word pairs in a particular corpus displaying the possible character substitutions up to a particular LD distance given a particular alphabet.

In (Reynaert et al., 2012) we have shown that TICCL is easily and unsupervisedly adaptable to other languages. It was shown to outperform VARD2 on historical spelling normalization of diachronic Portuguese text. The TICCLops web service currently has basic provisions for 18 languages. The lexicons have been derived from the available open-source dictionaries for Aspell[3].

[2]Available from GitHub at: `https://github.com/martinreynaert/TICCL`

[3]Aspell dictionaries: `ftp://ftp.gnu.org/gnu/aspell/dict/0index.html`

2.2. TICCL's prior results on the single book Gold Standard

We repeat the best results obtained in Reynaert (2014b) in Table 1. Best-first ranked, TICCL achieves an F-score of about 83%, the balanced combination of about 71% recall with an almost perfect precision of 99.8%. This raises the overall accuracy of the OCRed Gold Standard book from 88.94% to 94.51%, which is an admirable result for a fully automatic process. These best results were obtained by equipping the system with the best resources available, i.e. on the basis of the combined regular Dutch TICCL lexicon and the Dutch Institute for Lexicology or INL[4] historical Dutch lexicon, further enhanced with the INL historical names list.

The exceedingly high precision scores in our previous evaluation of the same TICCL run on the Early Dutch Books Online collection – now in Delpher[5] the books in the 'Boeken Basis' collection (E: 'Books Basic') prior to 1801 – left us wondering and prompted us to undertake the present new and qualitatively different evaluation which should shed more light on TICCL's performance on the whole collection, rather than on a single book.

rank	R	P	F
best-first ranked			
1	70.98	99.79	82.96
10 best-first ranked			
10	77.27	99.81	87.11

Table 1: Evaluation results on the task of fully automatically normalizing and OCR post-correcting as measured on the full DPO35 Gold Standard. TICCL corrected the book together with the rest of the 10,333 books EDBO collection on the basis of the combined regular Dutch TICCL lexicon and the INL historical Dutch lexicon further enhanced with the INL historical names list. R denotes recall, P precision and F the F-score. We first list the best-first ranked scores, then list results as measured on the 10 best-first ranked CCs.

2.3. Towards a new EDBO Gold Standard

We set out to randomly select and annotate 1,000 TICCL-corrected text strings (further: focus variants or short: variants) from the very same TICCL output file that gave the best results in the prior evaluation in Reynaert (2014b). EDBO is a very sizeable digital book collection as is evidenced by the statistics of its Dutch books in Table 2.

The TICCL output file from which we – through a small matching misshap – in fact drew 1,020 TICCL-corrected variants is 26GB and has 822,748,938 lines of CCs for a total of 9,027,945 focus variants. We estimate that at least 80% of the EDBO word types have in fact been created by OCR misrecognition. Note that all pairs of variants and CCs fall within LD 2 and that TICCL performed an exhaustive search within this Levenshtein distance limit. Our evaluation is limited to a very modest 0.011% sample. The set

[4] http://www.inl.nl/

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

Unit	amount
Books	10,333
Pages	1.7M
Tokens	435M
Types	20M

Table 2: Statistics on the Dutch books in EDBO

of 1,020 word strings selected nevertheless required about 50 hours for verification and annotation. Verification implies querying Delpher for the at first sight often unidentifiable text string. We found this was necessary for 66% of the focus variants in our sample. We discuss the format of our annotations in Section 3.5.

We annotated these 1,020 items according to the following guidelines. Per TICCL-corrected variant we annotated one CC. This was in principle the ranked CC that correctly resolves the variant. In the absence of a correct resolution, we annotated the best-first ranked CC.

In all, these 1,020 annotated TICCL-corrected variants yielded 90,240 Correction Candidates (CCs), or on average: 88.4 CCs per variant. This is well in line with the on average 91.3 CCs per variant in the full TICCL output. In production work, we would typically ask TICCL to return the top 3 or 5 or perhaps 10 CCs, but in this experimental setting we imposed no such restriction in order to be able to fully measure recall, however elevated the ranks of the CCs.

2.4. Word length versus numbers of CCs

length	# variants	# CCs	mean	median
6	127	34854	274.4	144
7	146	21731	148.8	88
8	143	10548	73.8	31
9	152	7948	52.3	20
10	126	4151	32.9	17
11	124	2792	22.5	9
12	75	1738	23.2	9
13	46	907	19.7	12
14	33	672	20.4	11
15	13	125	9.6	8
16	18	172	9.6	4
17	7	23	3.3	2
18	4	140	35.0	1
19	5	66	13.2	11
20	1	6	6.0	6

Table 3: Division of numbers of CCs retrieved by the system per word length in characters totalled for all variants of the particular length, followed by mean and median value per variant.

In Table 3 we study the division of numbers of CCs retrieved by the system per word length. Note we asked the system to work on words longer than 5 characters only. With some outliers, variants accrue less CCs both on av-

erage or according to the median the longer they are. This is in line with the findings by Choudhury et al. (2007) on other languages.
The variant 'Delcen' collected by far the most CCs, in all 2,669. It turned out to be OCR flotsam and jetsam. Remark that for this examination of CCs per word length we have so far not taken into account whether or not the variant was in fact corrected or indeed to be deemed correctable.
We intend to study whether there is a usable relation between elevated numbers of CCs, which signal a high level of confusion, and the correctability of the focus. A possible strategy may be to let the system back off from correction given a particular amount of CCs in so far that it is probably unlikely the ranking will provide the required best-first correction. This would be particularly true for short word forms. However, we know that some relatively frequent words may acquire an elevated amount of OCR variants.
When we study the outlier of length 18 in Table 3, the OCR-variant 'Desniettegenftaand' (Correct historical Dutch word form: 'Desniettegenstaand', E: 'Nevertheless'), we see that all except 1 of its 133 CCs are OCR-variants of the single correct form. This correct form only has corpus frequency 62, but is present in the historical lexicon. The summed corpus frequencies of its OCR-variants amount to 1,465 which means there are at the very least 23.6 times more variants in the EDBO corpus than the actual correct word form.

2.5. The new EDBO Gold Standard

The newly annotated sample represents a new Gold Standard for evaluation of post-correction on the EDBO, suitable for future evaluations of TICCL as well as of other post-correction systems. We make it publicly available[6].
When we were annotating we noticed we were getting hits on books from the 1930s. We knew the EDBO did not contain any books younger than the early 1800s. We soon learned the Delpher 'basic collection' of digitized works had recently been expanded by another 10,000 titles. Hits on the younger books we disregarded in the annotation, but we cannot rule out we did base our annotations on some works that had not in fact been part of the original set we post-corrected with TICCL. It is pitfalls like this that the Nederlab project (Brugman et al., 2016) is hoped to help prevent in the future, by allowing scholars to register exactly the collection of texts their researches were based on.
In the next section we re-evaluate TICCL's performance on EDBO.

3. TICCL on EDBO: re-evaluation

3.1. How we evaluated

We evaluate in terms of Recall, Precision and F-score according to van Rijsbergen (1975) and our own recommendations in Reynaert (2008).
We annotated 49 items or 4.8% as incorrectable OCR-junk. We still count them as False Negatives (FNs), i.e. incorrect variants for which TICCL failed to propose a correct version. Most seem caused by inadvertent OCR breakdown. These 62 items are nevertheless part of the accounting, even though we do not consider these items to be a real target for a post-correction system, that rather these should be solved by perhaps reOCRing the surrounding text snippet or by presenting them to human annotators in a crowd-sourcing environment. The fact remains that TICCL has tried to correct these – and necessarily failed, which is why are obliged to deal with them in our accounting.
In the accounting we consider both types of true target items of our post-correction effort, i.e. the True Positives or TPs: incorrect word variants that were properly corrected by TICCL and the other FNs, i.e. the non-OCR junk that TICCL also failed to correct.
The annotator for better or worse being the judge, correct versions in this context are both acceptable historical word forms, whether present in our historical lexicon or not, and acceptable contemporary forms, again whether present in the lexicon or not.
In order to get an idea about the precision of our system, we naturally also consider the False Positives (FPs), correct word forms that were incorrectly reported incorrect, i.e. for which TICCL returned CCs when it should have disregarded them.

3.2. Scores per rank

rank	R	P	F	TP	FP	FN	L	C
1	0.35	0.84	0.49	333	62	625	280	53
2	0.41	0.86	0.55	391	62	567	44	14
3	0.44	0.87	0.59	422	62	536	27	4
4	0.46	0.88	0.60	437	62	521	9	6
5	0.47	0.88	0.61	448	62	510	7	4
6	0.47	0.88	0.62	454	62	504	1	5
7	0.48	0.88	0.62	461	62	497	5	2
8	0.48	0.88	0.63	464	62	494		3
9	0.49	0.88	0.63	469	62	489	3	2
10	0.50	0.88	0.64	474	62	484	1	4
11	0.50	0.89	0.64	478	62	480	2	2
12	0.50	0.89	0.64	481	62	477	1	2
20	0.52	0.89	0.65	493	62	465	1	
31	0.52	0.89	0.66	498	62	460		1
54	0.52	0.89	0.66	502	62	456	1	
100	0.53	0.89	0.66	507	62	451	1	
205	0.53	0.89	0.67	508	62	450	1	
603	0.53	0.89	0.67	512	62	446	1	

Table 4: The evaluation is on word types. We present a selection of the scores per rank. A number of intermediate scores have been removed. R denotes recall, P precision and F the F-score. We further give the actual counts for True Positives (TP), False Positives (FP) and False Negatives (FN). The column labelled 'L' gives the counts for TPs present in the lexicon, the 'C' column those present in the corpus only.

We present evaluation scores as calculated from the 1,020 TICCL 'corrected' items in Table 4, i.e. the sum of TPs, FPs and FNs at any correction rank. The scores are presented per rank of the CCs.
In fact, since we pursue unsupervised fully-automatic OCR

[6]The annotated set is available from `http://ticclops.uvt.nl/LREC2016.EvaluationSet.txt`

post-correction, we should only be interested in the best-first ranking scores. However, since the objective of this evaluation is to study how TICCL might be improved and what the prioritization for the various possible system enhancements should be, we present a far wider range of results. These range up to rank 603, the highest at which we observed an actual 'correction'.

It is obvious that with 35% recall and 85% precision at rank 1 on the random sample – rank 1 is the only rank of true interest for fully automatic correction – TICCL performs far less well than the previous evaluation had led us to believe. We are forced to conclude that our first Gold Standard, the 1789 history book for children, presents a less representative sample of the full corpus than the new random Gold Standard.

In recall, at rank 10, we pass the fifty-fifty borderline between items successfully corrected and those that were not. This, in itself, shows we need to improve the ranking system. Precision scores naturally get higher with the rank, but the actual number of FPs remains unchanged at 62.

Of the TPs we list per rank whether the correction was due to the fact that the CC was present in the lexicon (column 'L') versus only in the corpus (column 'C'). Divided over historical versus contemporary word forms the division is: in the lexicon: 144 versus 250 word forms, in the corpus: 53 versus 65. This clearly shows the positive contribution of both lexicon and corpus derived word forms to the correction task.

3.3. Errors due to run-ons, splits and other language text

Of the FNs, at rank 603, we are left with 446 unsolved variants. Subtracting the 49 uncorrectable ones, we are left with 397 items for which the annotator saw a possible correction. We find that 110 of these are run-on words, i.e. two words concatenated through loss of the intermediate space. Next, there are 71 split words. Together, these amount to 181 items, or 17.75% of our random sample. To resolve these, TICCL will have to be equipped to be able to handle and correct word bigrams as we equipped our prior system TISC (Reynaert, 2005) to do. As a consequence, this is now high on our to-do list.

Non-Dutch text also accounts for a sizeable part of the randomly selected items, and of the system errors. We see that Latin text passages are responsible for 49 items, i.e. 23 FNs and 10 FPs. The remaining 16 were TPs, 5 of which in fact occur in our Dutch lexicon, the others corrected due to corpus occurrence. French text passages are responsible for 25 items, i.e. 13 FNs and 3 FPs. The remaining 9 were TPs, 5 of which occur in our Dutch lexicon, the other 4 were corrected due to corpus occurrence. We also found 4 German items, 2 FPs and 2 FNs. This problem might be alleviated by prior language identification per paragraph of text within the corpus which would then allow for non-Dutch paragraphs to be barred from correction (at least by the Dutch TICCL version). In our new corpus building workflow PICCL (Reynaert et al., 2015) we have the required tool in place for this. Another way forward may be to also provide Latin and French lexicons to the system.

3.4. The other errors

When we finally disregard the FNs due to run-ons or splits or non-Dutch text passages, we are left with 178 variants for which TICCL failed to produce a correct CC, or 17.45% of our random sample.

Based on the pretty comfortable results in recall in Reynaert (2014b) we initially on the basis of the new Gold Standard set out to explore in far more depth the ranking system in TICCL. We were to run an extensive series of ablation tests on the various ranking features to see which contribute most and what combinations perform best. In this we are waylaid, faced with the more disappointing recall in our current evaluation. We are now redirected to a more in-depth examination of what exactly prevents our system from better correction results on this so far unattributed large part of errors in particular.

We therefore next gauge how many more TPs we may expect to gain given that we further extend our system apart from implementing a solution for solving run-ons and splits, in case by finding a suitable algorithm for extending its reach in terms of LD while retaining sufficient precision. Remember that TICCL was asked to work within LD 2, i.e. to allow a divergence between variant and CC of 2 characters only. In Table 5 we list results for FNs, adding their numbers in LD bins. We discern between bin 1 for LDs 1 and 2, bin 2 for LDs 3 and 4, bin 3 for LDs 5 and higher.

To want to correct and properly rank LD 5 OCR-variants we deem overly ambitious. We are quite sure LD cases above 4 are not likely soon to be properly resolved by post-correction, so it seems from about 13 to 18% of the FNs will never become TPs by this route. Recall may nevertheless be more than doubled by solving run-ons and splits, extending TICCL's reach to LD 4 and better ranking. A lot can be gained, yet.

We calculated the LDs between the focus variants and respectively their historical correction and their contemporary correct counterparts. It can be seen that in our annotations we have tended to provide the contemporary correct form far more often than a historical form (of which there may be several, which might obfuscate results). However, the results also show that the LD from a historical variant to its historical correct form is often lower than to its contemporary correct counterpart. The percentages nevertheless indicate the same ranges.

LDs	# FNs	% FNs	total # FNs
historical variants			
1-2	54	38.30	141
3-4	69	48.94	141
5-	18	12.7	141
contemporary variants			
1-2	98	29.25	335
3-4	176	52.54	335
5-	61	18.21	335

Table 5: Binned correction results as counted for historical variants versus contemporary variants in the random Gold Standard.

3.5. Example and annotation file format

We focus on an example of an FN, representing many, that given we find a suitable and scalable algorithm we think we will be able to solve. This is a relatively long word, which typically returns fewer CCs, in this case, just 2. Note that in this TICCL experiment, we ran in case-sensitive mode, but case is disregarded in de LD calculation.

Vosfellaarten#1#vosfeftaarten#1#2#0.5@NS~V~
T: vossestaarten

Vosfellaarten#1#Vosfeftaarten#1#2#0.5

In the excerpt from our new gold standard above, the '#' delimited columns give: the OCR-variant, its corpus frequency, the CC, frequency of the CC, LD between OCR-variant and CC and lastly the ranking score. If applicable, the '@' delineates TICCL's output from our annotations, which are separated by '~', where column 1 gives our assessment – here 'NS' for 'not solved' and therefore an FN, column 2 has 'V' if we verified online at Delpher, column 3 'T': for contemporary or 'H:' for historical correct word form.

The LD between the variant and its incorrect CCs, indeed themselves OCR-variants, here is (disregarding capitalization): 2. The LD between the variant and its contemporary CC is: 4 (not shown). TICCL's ranking score confidence, here 0.5, is rather low. We have so far not studied the possible usability of this score towards correction – or perhaps backing off from correction – nor do we attempt this in the current paper. This we delegate to possible future work.

Note in the excerpt above that this hapax, i.e. this focus variant occurring just once in the whole EDBO, acquires 2 other hapaxes as CCs. The excerpt from the full TICCL output on EDBO below shows that these (disregarding capitalization) in turn, indeed as variants, are best-first corrected by the word form 'vossestaarten' (E: foxes' tails) occurring only in the lexicon (signalled by the (artificial) frequency '100,000,000' – frequencies below this signal words occurring only in the corpus, higher frequencies signal words present in both lexicon and corpus) and are next paired with other OCR-variants with corpus frequencies 4, 3, 36 and 4 respectively. So we see that there are in total at least 49 occurrences of this single word, for which not a single correct form is present in the corpus due to OCR misrecognition. Delpher indeed returns no hits on query 'vossenstaarten' in its Basic Book collection on books from before 1801.

vosfeftaarten#1#vossestaarten#100000000#2#0.863

vosfeftaarten#1#Vosfenftaarten#4#1#0.803419

vosfeftaarten#1#vosfenftaarten#3#1#0.794872

vosfeftaarten#1#Vosfeftaart#36#2#0.777778

vosfeftaarten#1#vosfeftaart#4#2#0.760684

As the image in Figure 1 shows, this book was printed in Fraktur, obvious primarily from the long 's', as indeed is the case with most books in the EDBO collection. The long 's' – mostly OCR misrecognized as 'f' – is a notorious problem, often seen by researchers as the most pressing one in digitized collections. It so far confounds our current solutions in TICCL, largely because the 's' is one of the most frequent characters used in Dutch. Words displaying this phenomenon are therefore still too often left uncorrected by our system as the f-s confusion coupled to another two other misrecognized characters are prevented from being corrected due to the LD limit set at 2.

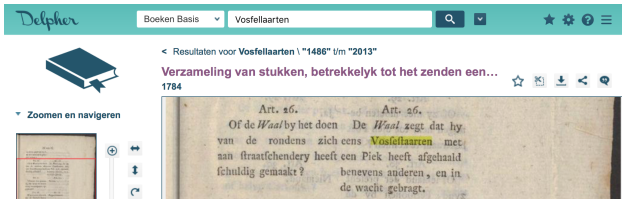

Figure 1: Image snippet from Delpher showing the query term and the search result highlighted in yellow on the page image (http://resolver.kb.nl/resolve?urn=dpo:7291:mpeg21:0453)

4. Discussion

4.1. Related gigascale correction work

TICCL can be run with LD 3, but this has a high processing cost and most probably results in lower precision. We demonstrate on the basis of the example we give in the previous section that in fact, the way it was run in Reynaert (2014b), it already did a great deal of the work required to have it correct up to LD 4, with the distinct possibility of retaining the better part of the precision achieved currently. The latter option seems the way to go.

The example given in the previous section is strongly reminiscent and in need of what Cucerzan and Brill (2004) call the iterative correction approach. On the basis of the incorrect queries users submit online, their approach iteratively searches for the correct solution of e.g. a query for 'anol scwartegger' – by intermediaries – into 'arnold schwarzenegger'. In fact, TICCL has at hand the necessary statistics to properly resolve 'Vosfellaarten' into 'vossestaarten' at run-time, using some adapted version of Viterbi search. An alternative solution may be found in the spelling system of Whitelaw et al. (2009) which follows a noisy channel model of spelling errors going back to Kernighan et al. (1990).

We need to more closely study and, if indeed applicable, emulate these possible solutions. Applicable in this context does mean tractable and making the most efficient use possible of data already available to TICCL. This is therefore not meant to be understood that we are about to toss out our own solutions, and replace them altogether with a noisy-channel or similar solution. This is to say that we now think both approaches deliver partial solutions and that the combination of our own anagram-hashing and corpus-statistics based solution together with a lightweight version of a Viterbi search or the noisy channel approach might well offer a far superior solution.

4.2. Methodological issues and differences between both EDBO evaluations

We here want to first focus on the cost of building evaluation sets for OCR post-correction. We move on to describe

the relative benefits of the full text versus the random sample evaluation methods.

Building the single EDBO book Gold Standard was undertaken in steps. First, at INL as a deliverable for the IMPACT European project, the book's OCRed text was manually transformed in an OCR ground truth. In CLARIN-NL project TICCLops, this ground truth was turned into a historical Gold Standard text. In the subsequent CLARIN-NL project @PhilosTEI we finally also added the contemporary Gold Standard text. We have previously written at length about the qualitative differences between the three versions (Reynaert, 2014a). The point we want to make here is that we estimate each step of this process to have cost about three weeks of work, i.e. nine weeks in all, say: 360 man-hours.

Collecting and annotating the random evaluation set cost us about 50 man-hours.

One phenomenon which cannot be inferred from the random sample evaluation method is that of 'letter spacing'. This is that instead of applying bold-face or italics to highlight a part of text, techniques which were often not an option for 18th century printers, they relied on inserting extra space between the individual letters of a word in order to lighten its appearance. More often than not, the OCR process renders this extra space as a full space mark, which in effect obliterates the words, very often names, in the frequency list TICCL derives from the corpus. Another deleterious effect of this letter spacing is that precisely the important information, which already the printers at the time drew attention to, is absent from the indexes of the online search systems and therefore irretrievable. In order to solve this problem, TICCL will need to be equipped with a special module geared at picking up elongated stretches of very short word tokens.

On the other hand, the full book evaluation told us nothing about the extent of the foreign language problem in our supposedly Dutch corpus. This book directed at children contains next to no text in other languages.

The random sample method being less costly, in conclusion, is better fit for providing a faster, overall view over the entire corpus. The full text method is no doubt recommended for obtaining a comprehensive view of the system's capabilities, including what the system is capable of doing for short words, with the caveat that if important features of the corpus happen not to be present in the full text, these will be overlooked.

4.3. An appeal for better diachronic lexicons and name lists

Whitelaw et al. (2009) further advocate that given sufficiently large corpora, one can in fact forego using a validated lexicon and induce one from the corpus. In fact, we already proclaimed the same for spelling correction in (Reynaert, 2005). A conclusion to be drawn from our current re-evaluation is that for OCR post-correction the strategy is not advisable. The examples and attendant frequencies we have presented show all too clearly that with very noisy OCR corpora the frequencies of misrecognized OCR variants may easily outweigh those of the correct word forms. We advocate that far more work be done on building large-coverage, validated diachronic lexicons and name lists for Europe's languages.

4.4. Discussion of the long 's' problem

We next discuss the long 's' problem in books printed in Fraktur. The very same problem in respect to the same corpus was discussed earlier by de Does and Depuydt (2013) who mainly sought to improve the OCR process itself by means of the INL historical lexicon and name list mentioned before. These too were deliverables of the European project IMPACT and are available through the Impact Centre of Competence[7].

The famous Fraktur long 's' is mostly rendered as 'f' by the OCR process. Our prior example of 'vossestaarten' already showed that confusion of the long 's' by 'f' is not necessarily the case. However, it is the feature of digitized diachronic texts most often commented on by researchers working on digitized text (cf. De Does,). We therefore here analyse the phenomenon more in depth.

In fact, we find it to be the only 34th most common character substitution in the very large corpus we happen to work on. Of the 706,947 unique f-s variants retrieved by TICCL, we count 437,923 word pairs sporting only the f-s confusion, with another 269,024 showing this in their bag-of-characters, accompanied by further character transpositions. An example of the latter is the variant 'Blixfem' with CC 'Blixems' with LD 2 (contemporary: 'bliksem' versus 'bliksems', i.e. the singular and plural forms for 'lightning').

The strength of the powerful anagram hashing variant retrieval mechanism may in this respect be seen to be a weakness in that it may just as easily pair two further unrelated words. We try to correct for this in the ranking of the CCs by checking whether the first characters of the word pair match and/or whether the final two characters match. This actually fails in the pair 'geflooten' (E: whistled) with corpus frequency 10,183 and 'geslooten' (E: closed) with lexicon/corpus frequency 100,000,101, i.e. 100M for being present in the validated lexicon and 101 actual corpus occurrences. The historical word form for 'gefloten' should by rights be in the lexicon, but it is not. If it had been, this pair would have been blocked, i.e. we do not try to solve confusables or real-word errors at this time. It is very likely that in fact the major bulk of the occurrences of 'geflooten' should be corrected as 'geslooten'. Or 'gesloten' if we attempt to modernize the text, since the double vowels in open syllables were abandoned in the Dutch spelling reforms around 1950.

Admittedly, f-s confusion represents a major problem. The top 8 most frequently OCR-misrecognized word types that were printed with a long s and therefore were misrecognized as having an f, already account for over one million or 0,25% of the word tokens in our corpus. We list them in Table 6. One of these was not corrected at all, in spite of its 130 CCs. The string is an f-s confused variant for the highly frequent diachronic abbreviation 'voorsz.' which stands for a range of attested forms for 'voorzegd' (E: foresaid). For further use of TICCL we should simply add this to the lexicon, for EDBO we might now apply 'absolute correction'

[7] http://www.digitisation.eu/

in the sense of Pollock and Zamora (1983): once a particular incorrect and unambiguous word form has been attested, one may indeed simply replace it by its correct form whenever encountered.

variant	frequency	best-first CC	English
eerfte	218,932	eerste	first
voorfz	148,103	voorts *	further
fchoon	131,667	schoon	clean
menfchen	130,727	menschen	people
Misfive	121,293	missive	letter
Commisfie	119,786	commissie	commission
ftellen	104,222	stellen	to put
tusfchen	101,008	tusschen	between
# Tokens	1,075,738		

Table 6: Top 8 f-s confusions corrected by TICCL with corpus frequencies and (lower-cased) best-first ranked CCs. Together these 8 word types would account for over 1 million or 0,25% corrected tokens in EDBO if the asterisk marked one had also been properly corrected.

4.5. Confusables or real-word errors

The top character confusions observed in the TICCL output list may be responsible for any number of confusables. Given the highly frequent word 'heeren' (E: lords, gentlemen) and the higly frequent h-b confusion we encounter far more 'beeren' (E: bears or boars) in the texts than is right. Likewise with the e-o confusion which makes especially the KB newspaper collections teem with 'hoeren' (E: whores). The combination of both confusions then returns gentlemen to the status they originally often had, i.e. that of 'boeren' (E: farmers).

Add to this that in Dutch compounds are written as single words and another layer of complexity unfolds. We find 'tempelhoeren' 4 times in EDBO. Consultation of the originals in Delpher shows that twice the printed word is in fact 'temple whores' (cf. Figure 2), but even so twice 'temple lords' should have been recognized (cf. Figure 3). The converse may be true for any number of the 126 occurrences of 'Tempelheeren' (disregarding capitalisation).

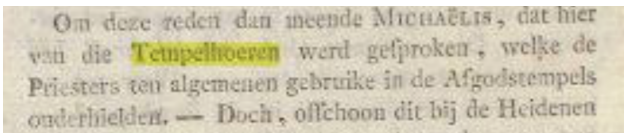
Om deze reden dan meende Michaëlis, dat hier van die Tempelhoeren werd gesproken, welke de Priesters ten algemenen gebruike in de Afgodstempels onderhielden. — Doch, offchoon dit bij de Heidenen

Figure 2: Delpher image snippet showing the correct search result 'Tempelhoeren' (i.e. Temple whores) for the query term 'Tempelhoeren' highlighted in yellow. (http://resolver.kb.nl/resolve?urn=dpo:10706:mpeg21:0106)

This situation is even further compounded by the fact that in most lexicons or dictionaries, due to their immense productivity, most compounds whose meaning can readily be inferred from the composing words are simply not listed. We defer possible correction work on confusables in OCRed texts to the time when we manage to solve the non-word errors with sufficient accuracy.

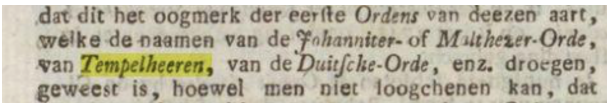
dat dit het oogmerk der eerfte Ordens van deezen aart, welke de naamen van de Johanniter- of Malthezer-Orde, van Tempelheeren, van de Duitfche-Orde, enz. droegen, geweest is, hoewel men niet loogchenen kan, dat

Figure 3: Delpher image snippet showing the divergent search result 'Tempelheeren' (i.e. Temple lords) for the query term 'Tempelhoeren' highlighted in yellow. (http://resolver.kb.nl/resolve?urn=dpo:1184:mpeg21:0103)

4.6. Possible uses for TICCL corrected ranked lists

We discern three possible viable and productive uses for TICCL ranked correction lists.

First, the variants linked to the best-first ranked CCs given a user's query might be added to the query to enhance retrieval. We should recommend to the National Dutch Library or KB that this facility be provided within Delpher as an option to the user. This would be analogous to the option already available where a user's query in contemporary Dutch is expanded with the known historical word forms available from INL through a RESTful web service[8].

Second, given a suitable user interface on a text collection, the user might be enlisted in a crowdsourcing setting to help identify the correct CC on the basis of the top n CCs being presented to him besides the actual text image. We hope to develop this in the framework of the CLARIAH project PICCL, currently underway.

Third, the actual texts may be edited, i.e. effectively corrected by automatically replacing variants with the best-first ranked CC.

This last is the use we put these to in the Nederlab project. We there copy each text paragraph, identify this as the TICCL-corrected one with a suitable XML attribute, correct it and store it alongside the original OCR version. In this we aim not only to post-correct the OCRed text but also to modernize the Dutch. This is in order to subsequently have the text further linguistically enriched, i.e. have it tokenized, sentence split, lemmatized, Part-of-Speech tagged and labelled for Named Entities. Besides allowing for corrected paragraphs, the FoLiA XML format (van Gompel and Reynaert, 2013) we use offers the possibility of also incorporating the top n ranked CCs.

5. Concluding remarks

It is sobering to determine that a great deal of the tangible potential of our system is not currently realised. On the basis of our new gold standard, we have demonstrated that TICCL currently in a single LD 2 run manages to successfully pair – best-first ranked: just about one third, regardless of rank: just about half – of the OCR variant types with an acceptable historical or contemporary correct word form. At the same cost in terms of processing time and computer cycles spent, based on our sample, it pairs so far

[8] http://sk.taalbanknederlands.inl.nl/LexiconService/

countless OCR-variants outside the LD limit to these correctly linked OCR-variants that are within the LD limit. It brings these further-off OCR-variants 'half way home'. If we find a way of bringing these home all the way, solve run-ons and splits as well and manage to improve its ranking, TICCL will effectively make a tangible difference to our digitised text legacy by noticeably raising overall text accuracy.

6. Acknowledgments

Martin Reynaert acknowledges being funded by the Netherlands Organisation for Scientific Research in NWO 'Groot' project Nederlab and CLARIAH WP2 and WP3 project PICCL.

7. Bibliographical References

Brugman, H., Reynaert, M., van der Sijs, N., van Stipriaan, R., Tjong Kim Sang, E., and van den Bosch, A. (2016). Nederlab: Towards a Single Portal and Research Environment for Diachronic Dutch Text Corpora. In *Proceedings of the Tenth International Language Resources and Evaluation Conference (LREC-2016)*, Portorož, Slovenia. ELRA.

Choudhury, M., Thomas, M., Mukherjee, A., Basu, A., and Ganguly, N. (2007). How difficult is it to develop a perfect spell-checker? A cross-linguistic analysis through complex network approach. *Proceedings of the second workshop on TextGraphs: Graph-based algorithms for natural language processing*, pages 81–88.

Cucerzan, S. and Brill, E. (2004). Spelling correction as an iterative process that exploits the collective knowledge of web users. In Dekang Lin et al., editors, *Proceedings of EMNLP 2004*, pages 293–300, Barcelona, Spain, July. Association for Computational Linguistics.

de Does, J. and Depuydt, K. (2013). Lexicon-supported OCR of eighteenth century Dutch books: a case study. In Richard Zanibbi et al., editors, *DRR*, volume 8658 of *SPIE Proceedings*. SPIE.

Kernighan, M. D., Church, K. W., and Gale, W. A. (1990). A spelling correction program based on a noisy channel model. In *COLING-90*, volume II, pages 205–211, Helsinki.

Pollock, J. and Zamora, A. (1983). Collection and characterization of spelling errors in scientific and scholarly text. *Journal of the American Society for Information Science*, 34(1):51–58, January.

Reynaert, M., Hendrickx, I., and Marquilhas, R. (2012). Historical spelling normalization. A comparison of two statistical methods: TICCL and VARD2. In Francesco Mambrini, et al., editors, *Proceedings of ACRH-2*, pages 87–98. Lisbon: Colibri.

Reynaert, M., van Gompel, M., van der Sloot, K., and van den Bosch, A. (2015). PICCL: Philosophical Integrator of Computational and Corpus Libraries. In *Proceedings of CLARIN Annual Conference 2015 – Book of Abstracts*, pages 75–79, Wrocław, Poland. CLARIN ERIC.

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

Reynaert, M. (2008). 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. ELRA.

Reynaert, M. (2010). Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal on Document Analysis and Recognition*, 14:173–187.

Reynaert, M. (2014a). On OCR ground truths and OCR post-correction gold standards, tools and formats. In A. Antonacopoulos et al., editors, *Proceedings of Digital Access to Textual Cultural Heritage, Datech 2014 Conference, Biblioteca National de España, Madrid*, pages 159–166, New York, NY, USA. ACM.

Reynaert, M. (2014b). Synergy of Nederlab and @PhilosTEI: diachronic and multilingual Text-Induced Corpus Clean-up. In *Proceedings of the Ninth International Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland. ELRA.

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

van Rijsbergen, C. J. (1975). *Information Retrieval*. Butterworths, London.

Whitelaw, C., Hutchinson, B., Chung, G., and Ellis, G. (2009). Using the web for language independent spellchecking and autocorrection. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*, EMNLP '09, pages 890–899, Stroudsburg, PA, USA. Association for Computational Linguistics.