OCR of Historical Printings of Latin Texts: Problems, Prospects, Progress

Uwe Springmann[*]
CIS
University of Munich
springmann@cis.uni-muenchen.de

Dietmar Najock[†]
Institute of Greek and Latin Languages and Literatures
Freie Universität Berlin

Hermann Morgenroth[‡]
Institute of Greek and Latin Languages and Literatures
Freie Universität Berlin

Helmut Schmid
CIS
University of Munich
schmid@cis.uni-muenchen.de

Annette Gotscharek[§]
CIS
University of Munich
annettegotscharek@t-online.de

Florian Fink
CIS
University of Munich
finkf@cis.uni-muenchen.de

ABSTRACT

This paper deals with the application of OCR methods to historical printings of Latin texts. Whereas the problem of recognizing historical printings of modern languages has been the subject of the IMPACT program[1], Latin has not yet been given any serious consideration despite the fact that it dominated literature production in Europe up to the 17th century. Using finite state tools and methods developed during the IMPACT program we show that efficent batch-oriented post-correction can work for Latin as well, and that a lexicon of historical Latin spelling variants can be constructed to aid in the correction phase. Initial experiments for the OCR engines Tesseract and OCRopus show that some training on historical fonts and the application of lexical resources raise character accuracies beyond those of Finereader and that accuracies above 90% may be expected even for 16th century material.

Categories and Subject Descriptors

I.7.5 [**Document and Text Processing**]: Optical character recognition (OCR)—*Latin language, historical documents, recurrent neural networks*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Latin language*

[*]Corresponding author.

[†]Send email to klassphi@zedat.fu-berlin.de, attn. Ms. Regina Davis.

[‡]Author has left FU Berlin.

[§]Author has left CIS.

[1]`http://www.digitisation.eu/`

DATeCH 2014 May 19 - 20 2014, Madrid, Spain
 ACM 978-1-4503-2588-2/14/05...$15.00.
http://dx.doi.org/10.1145/2595188.2595205

.

1. INTRODUCTION

The currently renewed awareness of the European cultural heritage embodied in its literary sources has produced research and tools for tackling the problem of making historical full texts available: How can we significantly lower the cost barrier (in terms of both time and money) to convert scanned page images into searchable full text? This question has led to adaptations of existing tools in the area of optical character recognition (OCR) to account for special features by which the page images of historical texts differ from modern book pages: historical fonts including ligatures, historical spelling variants, somewhat displaced characters (resulting from historical printing processes), fuzzy character boundaries due to ink creep into the paper over time, paper degradation resulting in dark backgrounds, blotches, cracks, dirt, and bleed through from the following page. While most of these features are characteristic of any surviving historical documents, fonts and spelling are strongly language dependent and need special attention if the accuracy resulting from the OCR process is deemed insufficient for a specific research interest. To distinguish between historical spellings (which we might choose to change to their modern form or rather leave in their historically correct printed form) and true OCR errors needing correction, dictionaries of historical spellings have been produced for a variety of modern languages such as German, French, English, Italian and Spanish. If we choose to leave the historical printed spelling in our electronic text version, these lexica of historical spellings (both of modern word forms and their lemmata) will also be needed for later information retrieval (IR): searching for a modern lemma should turn up all historical variants in the text as well.

However, to our knowledge nobody has yet successfully tackled the OCR problem for historical printings of Latin texts, although Latin has been the predominant language for documents in Europe up to the 17th century ([7], p. 5) and although the common Latin tradition is one of the defining features of Europe´s cultural identity. Whatever the reason for this gap, we have set out to remedy it.

To sum up, in order to make historical printings of Latin texts available as electronic text data as opposed to page images, we need an OCR system capable of recognizing historical fonts including ſ (long s), ligatures Æ, æ, Œ, œ, ﬅ,

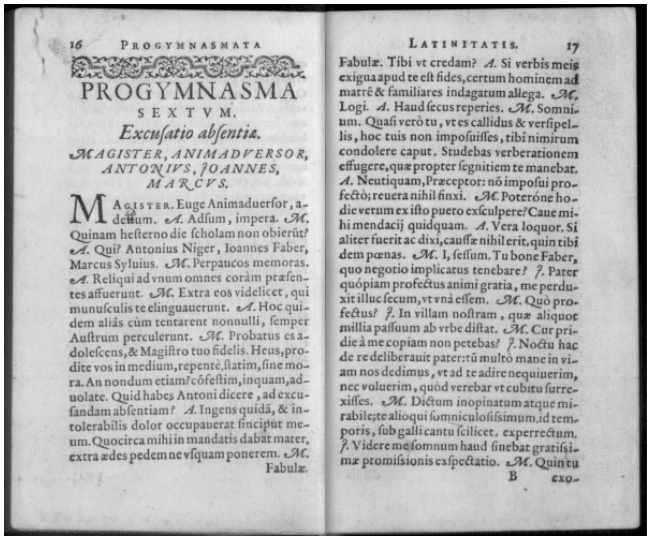
sometimes also ſt, ct, and the use of diacritic marks in the humanist system for disambiguation (e.g. *cùm* with grave accent is a conjunction, *cum* without accent is a preposition; *feminâ* with circumflex marking long vowel is ablative, *femina* is nominative). Historical spelling variants must be taken into account in order to distinguish between incorrect (OCR recognition error) and historically correct spellings (e.g. classical *femina* is written *fœmina* etc.). For IR purposes, we must make sure that a search for a string in modern spelling will be found regardless of the historical variants. Ideally, the OCR should also be able to automatically recognize the frequent Greek quotations appearing in Latin texts printed with their own extensive set of ligatures having long since fallen into disuse. Otherwise, one can choose to insert the Greek passages by hand later in the correction phase. In this paper we record first steps to achieve these goals, i.e. our efforts to build lexical resources for efficient postcorrection (including a lexicon of historical spelling variants).

The remainder is organized as follows: In Sect. 2 we have a look at the current status when applying available OCR methods to historical Latin printings, Sect. 3 gives information about our Latin lexicon of word forms, Sect. 4 looks briefly into the spelling variants and how to cope with them and Sect. 5 and 6 report the results of some training for Tesseract and OCRopus which look very promising for our goals. Sect. 7 describes the conclusions and the next steps we are taking.

2. CURRENT STATUS

The current state of the art in OCR for Latin historical printings is characterized by little more than the application of current OCR engines such as ABBYY Finereader[2], BIT Alpha[3] and the open source alternatives Tesseract[4] and OCRopus[5]. Typically one of these engines is applied to scanned page images and a human editor corrects the OCR output afterwards by inspecting the original image. While this works reasonably well for modern printings, for historical printings up to the 19th century the amount of necessary correction work makes it often more worthwile to transcribe the original directly, which has been the method of choice for any large scale digitization efforts. Some efforts to OCR early modern Latin texts have been reported using the open source framework Gamera[6] [11] where it was concluded that without lexical resources a character accuracy of 80% was achieved with Gamera against 84% with Finereader.

Later postcorrection depends on a lexicon of word forms with good coverage and a lexicon of historical spelling variants. The latter resource does not yet exist. For the former, the words[7] program of the late William Whitaker exists, giving roughly 1 million word forms derived from about 39,000 lemmata, as well as the Perseus[8] resource of about 240,000 classically attested word forms with 49,000 lemmata coming from the analysis of Perseus texts. A third alternative consists in using the spell checking lexicon for OpenOffice/LibreOffice constructed by Karl Zeiler[9]. The Perseus resources have been used in the OCR work on Greek critical editions of the 19th and 20th century with their frequent Latin parts [1].


16 PROGYMNASMATA

PROGYMNASMA
SEXTVM.
Excuſatio abſentiæ.
MAGISTER, ANIMADVERSOR, ANTONIVS, IOANNES, MARCVS.

MAGISTER. Euge Animaduerſor, adeſſum. *A.* Adſum, impera. *M.* Quinam heſterno die ſcholam non obierût? *A.* Qui? Antonius Niger, Ioannes Faber, Marcus Syluius. *M.* Perpaucos memoras. *A.* Reliqui ad vnum omnes coràm præſentes affuerunt. *M.* Extra eos videlicet, qui munuſculis te elinguauerunt. *A.* Hoc quidem aliàs cùm tentarent nonnulli, ſemper Auſtrum percuſerunt. *M.* Probatus es adoleſcens, & Magiſtro tuo fidelis. Heus, prodite vos in medium, repentè, ſtatim, ſine mora. An nondum etiam? cõſeſtim, inquam, adsuolate. Quid habes Antoni dicere, ad excuſandam abſentiam? *A.* Ingens quidã, & intolerabilis dolor occupauerat ſinciput meum. Quocirca mihi in mandatis dabat mater, extrà ædes pedem ne vſquam ponerem. *M.* Fabulæ.

LATINITATIS. 17

Fabulæ. Tibi vt credam? *A.* Si verbis meis exiguæ apud te eſt fides, certum hominem ad matrê & familiares indagatum allega. *M.* Logi. *A.* Haud ſecus reperies. *M.* Somnium. Quaſi verò tu, vt es callidus & verſipellis, hoc tuis non impoſuiſſes, tibi nimirum condolere caput. Studebas verberationem effugere, quæ propter ſegnitiem te manebat. *A.* Neutiquam, Præceptor: nõ impoſui profectò; reuera nihil finxi. *M.* Poteróne hodie verum ex iſto puero exſculpere? Caue mihi mendacij quidquam. *A.* Vera loquor. Si aliter fuerit ac dixi, cauſſæ nihil erit, quin tibi dem pœnas. *M.* I, feſſum. Tu bone Faber, quo negotio implicatus tenebare? *F.* Pater quópiam profectus animi gratia, me perduxit illuc ſecum, vt vnà eſſem. *M.* Quò profectus? *F.* In villam noſtram, quæ aliquot millia paſſuum ab vrbe diſtat. *M.* Cur pridie à me copiam non petebas? *F.* Noctu hac de re deliberauit pater: tũ multò mane in viam nos dedimus, vt ad te adire nequiuerim, nec voluerim, quòd verebar vt cubitu ſurrexiſſes. *M.* Dictum inopinatum atque mirabile; te alioqui ſomniculoſiſſimum, id temporis, ſub galli cantu ſcilicet, experrectum. *F.* Videre me ſomnum haud ſinebat gratiſiſimæ promiſſionis exſpectatio. *M.* Quin tu

B exo-

Figure 1: Page 16 and 17 of Pontanus (1589)

Table 1: Character accuracies in % for sample pages

Page	Finereader 11	Tesseract 3.03	OCRopus 0.7
15	87.79	80.88	80.70
16	82.94	77.41	76.94
17	85.25	75.98	86.07
18	85.93	79.51	85.53
19	87.94	80.09	79.09

In a recent study by the Berlin State Library ([3], p. 128) BIT Alpha was found to be too complex to be used in-house and should therefore be left as a service to the company (except for the training of fonts); Tesseract and OCRopus lack adequate documentation, but in the case of Tesseract this is partly remedied by an active community.

In order to get an idea of what character accuracies one can expect with these engines, we give in Table 1 the accuracies for some 300 dpi scans of a 16th century book[10] ([9], Fig. 1), Pontanus' *Progymnasmata Latinitatis*, printed 1589 in Ingolstadt. The results have been obtained with default options and without any training. Accuracies have been measured with the OCR evaluation tools from ISRI/UNLV written by Stephen Rice and Thomas Nartker and adapted for UTF-8 by Nick White[11].

Finereader gets the best results, but because it cannot be trained on fonts (only single character training is available) there is unfortunately no room for improvement for an end user. The results of OCRopus are also very promising, as they are comparable to Abbyy for pp. 17 and 18, but they

[2] http://www.abbyy.com
[3] http://bit.dyndns.biz
[4] http://code.google.com/p/tesseract-ocr/
[5] http://code.google.com/p/ocropus/
[6] http://gamera.informatik.hsnr.de/addons/ocr4gamera/
[7] http://archives.nd.edu/whitaker/words.htm
[8] http://www.perseus.tufts.edu/hopper/opensource/downloads/texts/hopper-texts-GreekRoman.tar.gz
[9] http://extensions.openoffice.org/de/project/latin-spelling-and-hyphenation-dictionaries
[10] http://nbn-resolving.de/urn:nbn:de:bvb:12-bsb00037616-2
[11] https://gitorious.org/ancient-greek-training-for-tesseract/ocr-evaluation-tools

show more variation. This is due to its present inability to sort out floral decor often found in books of this period, so it tries to recognize characters where there are none (page 15, 16 and 19). The main error sources for both Finereader and OCRopus are f (long s) misrecognized as f and the difficulty to find word boundaries. The first problem should be addressable by training on fonts and the second by the use of lexical resources, at least in the postcorrection phase. Without font-specific training, Tesseract has markedly lower accuracies and inspection of its errors shows that the top error source are e's misrecognized as c's, an error which shows up less frequently in the other two engines (in page 17, there are just 4 cases in the OCRopus output).

On the performance of BIT Alpha, Thomas Stäcker reported ([3], p. 130) that in the recognition of 17th and 18th century printings with mixed fonts (e.g. Antiqua and Italics for Latin and Gothic (Fraktur) for German) a mean character accuracy of 96% has been achieved after extensive training and that there seems to be an upper limit of 99% on single pages for pattern recognition alone. He assumes that this is due to the frequent misrecognition of badly printed e's as c's. Therefore lexical resources will still be necessary if one wants to go beyond high levels such as 99%, which still means 12-20 errors per page or one error every other line.

We conclude that in order to get to higher accuracy we can follow two strategies: Training on the characteristic fonts for the material of interest and applying lexical resources. ABBYY has done some work during the IMPACT program on both Gothic scripts and historical lexica for some modern languages, but these developments were a company decision and further adaptations cannot be done by individual users. This leaves us with Tesseract and OCRopus as the only viable options for an academic research environment. Both have successfully been trained to Gothic scripts so there is hope they can be adapted to historical fonts as well. The second strategy, the application of lexical resources, will be covered first, in the next two sections.

3. LEXICON OF WORD FORMS

One of the outcomes of IMPACT was the postcorrection tool PoCoTo based upon the profiling technology described in [12]. The tool enables semi-automatic batch correction by drawing statistical inferences about error series. These inferences allow specific correction candidates to be proposed which have the highest probability as derived from the document specific error profile (for the details see [15]). The profiler requires a lexicon of word forms and a list of patterns describing the difference of historical spelling variants from their modern counterparts together with their initial probability, taken from a background corpus. In the case of early modern Latin, typical patterns are the letters *j* and *y* that have since been normalized to *i*, the variation between double and single consonants (*imo* for modern *immo*), forms such as *quum* for modern *cum* etc.

The profiler, which is realized as a minimal deterministic Levenshtein automaton, needs a list of word forms as complete as possible. There are two possible ways by which such a list can be created: Either (A) by building a big corpus of existing electronic texts and tokenizing the attested forms, or (B) by starting from a lexicon of ground forms or lemmata, from which all possible word forms are generated. The first path has been taken by the Perseus project resulting in about 240,000 attested word forms from 49,000 lemmata. However,

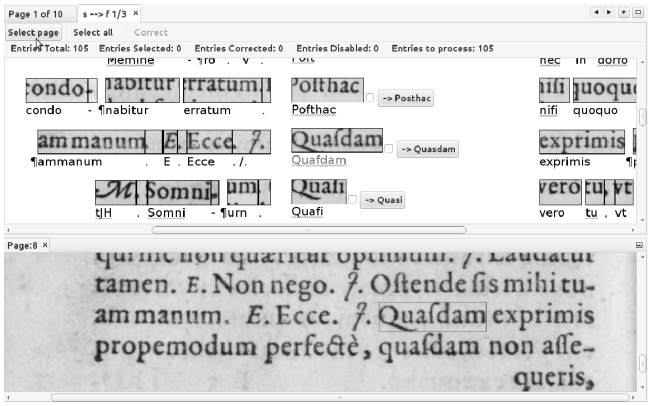


Figure 2: The postcorrection tool PoCoTo with an error series for f (long s) confused with f.

this leaves out the majority of about 2 millions forms that could be generated from such lemmata and that will show up eventually when treating large amounts of postclassical texts. Therefore we chose to start from a lemma lexicon which had been compiled in the 1980s in the group of Dietmar Najock at Freie Universität Berlin. This lexicon of about 70.000 lemmata is built mainly from the entries of Georges' Handwörterbuch [4] with additional proper names from Lewis & Short [8]. The lemmata were then expanded into full forms by a program written by Hermann Morgenroth ([10] p. IX-XII). The resulting file consisting of about 2 million full forms with their corresponding lemma and morphosyntactic annotation was used for the lemmatization of ancient Latin texts in order to build concordances. When new words showed up in the course of this work they were incorporated into the lexicon.

Since the lexicon and the program generating full forms and the morphosyntactic annotation had lain dormant for some years we decided to build in addition a new morphology based upon the Stuttgart Finite State Tools[12] (SFST) authored by Helmut Schmid [13]. The new transducer based morphology is not restricted to any specific inventory of lemmata and can derive new and classically unattested word forms by modeling the morphological processes of derivation and compounding.

While this work has not yet been completed we have been able to build a prototype profiler to show that the batch correction of historical printings of Latin texts is possible even when spelling variants must be taken into account (see Fig. 2). The postcorrection tool requires word coordinates in either ABBYY xml or hocr format, so currently only Finereader and Tesseract are supported.

4. HISTORICAL SPELLING VARIANTS

Further progress can be made in recognition as well as postcorrection if a lexicon of known historical spelling variants is available. In the case of early modern Latin the situation is less dramatic than for, e.g., early German, where lack of any orthographic standard got the same word spelled differently in the same document. For Latin a historical orthography was used which changed slowly over time. At CIS we have the "Lextractor" tool [5] available for the purpose of building

[12] `http://www.cis.uni-muenchen.de/~schmid/tools/SFST/`

up a lexicon of historical spelling variants (see Fig. 3). When we get hold of some diplomatically transcribed texts (not normalized to modern spelling) we will be able to expand our historical spelling lexicon. This lexicon will also be a source of out-of-dictionary words whenever a word is not a spelling variant and does not contain an error. A further advantage will arise for the use of IR: A query for a modern lemma in a historical document will find attested spelling variants as well.

5. TRAINING OF TESSERACT

Apart from applying lexical resources to get better results, one can train an OCR engine to the peculiarities of the historical fonts (as mentioned in Sect. 2). Two general routes can be taken: Either the training is done on real printed images, which means that a human user must prepare the ground truth closely corresponding to the page images including printing errors, left out spaces and other ideosyncrasies. There is no need, however, to record any special printed letter forms (glyphs) such as ligatures, as long as any single printed glyph corresponds to a definite modern character or a combination of characters. This is manual, tedious, error prone and therefore expensive work.

The other training route consists in generating artificial images from available texts serving as ground truth, employing historical fonts. To make the images look more closely like an old printing, they should be filtered and degenerated in their quality, resulting in somewhat fuzzy and jittery images similar to their scanned counterparts. Here the problem is that one must generate all the special glyphs one expects to find in the images. However, some ligatures and other glyphs are not available in Unicode, and the Unicode Consortium has declared that further ligatures will not be encoded[13] as they do not belong to the characters, but instead to their specific font dependent presentation form. Therefore one is left with fonts encoding these glyphs in the Private Use Area (PUA) left by Unicode for user specific additions. However, different fonts treat the PUA code points differently, so that one cannot generally produce page images from the same ground truth with different fonts.

For our training experiments we used a digital reproduction of the so-called Fell types, a collection of 17th century fonts bequeathed by John Fell, Bishop of Oxford, to the University at his death in 1686, of which 15 fonts were made available in digital form by Igino Marini[14]. Using the Fell types one can both map the historical feature set to some extent and get some variations in font shapes.

Starting from an abitrary training text containing all the glyphs one wants to train several times (a few text pages are therefore sufficient), the new training tools of Tesseract 3.03 allow the automatic image and box file generation (command text2image). Training on 5 regular and 5 italic fonts takes only a few minutes. The results of this training are shown in Table 2 under column heading "Tess. (font)" and show an increase in accuracy between 6 and 12 percentage points compared to Table 1.

To get an estimate of the potential improvements due to lexical resources we provided Tesseract with the "ideal lexicon" containing all word forms present in the 5 pages. This is a simulation of a lexicon with 100% coverage representing the ideal limit. Another 3 to 5 percentage points could ideally be gained, which would raise the accuracy above 90% in 4 of 5 pages (column "Tess. (font+lex)").

[13] `http://www.unicode.org/faq/ligature_digraph.html`
[14] `http://iginomarini.com/fell/the-revival-fonts/`

Table 2: Character accuracies after some training

Page	Tess. (font)	Tess. (font+lex)	OCRopus (font)
15	91.02	93.90	83.66
16	80.12	85.65	78.00
17	85.41	91.56	86.80
18	88.29	92.68	89.59
19	86.06	90.15	80.97

The experiments with font training have by no means been systematic and thorough, but they show both the importance and the potential for training on historical fonts.

6. TRAINING OF OCROPUS

OCRopus (0.7), formerly a front-end for Tesseract, employs a new recognizer based upon the Long Short Term Memory (LSTM) architecture of recurrent neural networks (RNNs) originated by Hochreiter & Schmidhuber [6] and recently shown to outperform any other algorithm in image recognition competitions. OCRopus 0.7 shows excellent results even without any language model (which cannot currently be added to it) with a character error rate of about 1% for modern scripts ([2], [14]).

As in the case of Tesseract, one can either train on real images if ground truth is available or generate artificial images using historical fonts resembling the printed documents of interest. We used the same training text and Fell types as with Tesseract. The training proceeds by feeding the neural network repeatedly single text lines together with their ground truths. The number of lines randomly selected from a training text of several thousand text lines needed to get the cited error rates of 1% or better ranges typically from 100,000 to 1 million. For our experiment we just trained on the stock of 400 training text lines out of which 44,500 random selections were drawn. The training effort in terms of CPU cycles and therefore computing time is substantially higher than for Tesseract: at about 1s/line, the training took about 12 hours.

Already with this modest amount of training the problem of long s and its ligatures was solved. The remaining errors are unrecognized inter-word spaces and bad line segmentation in the presence of images and floral decorations. Pages 17 and 18 now have accuracies better than those of ABBYY by 1 to 3.5 percentage points. The references cited above with character error rates of about 1% give hope that more training can achieve much better accuracies still.

7. CONCLUSIONS AND FUTURE WORK

Whereas current OCR methods work admirably well even without Latin language support in the case of defect-free printings of modern fonts and high signal to noise ratios, i.e. uniformly black characters on a uniformly white background, they fail badly already for degraded character shapes, and get much worse when historical font features and higher noise levels must be taken into account. We have shown that we are able to build a statistical profiler for Latin which enables batch correction on statistically inferred error series and that we can also build a lexicon of historical spellings which

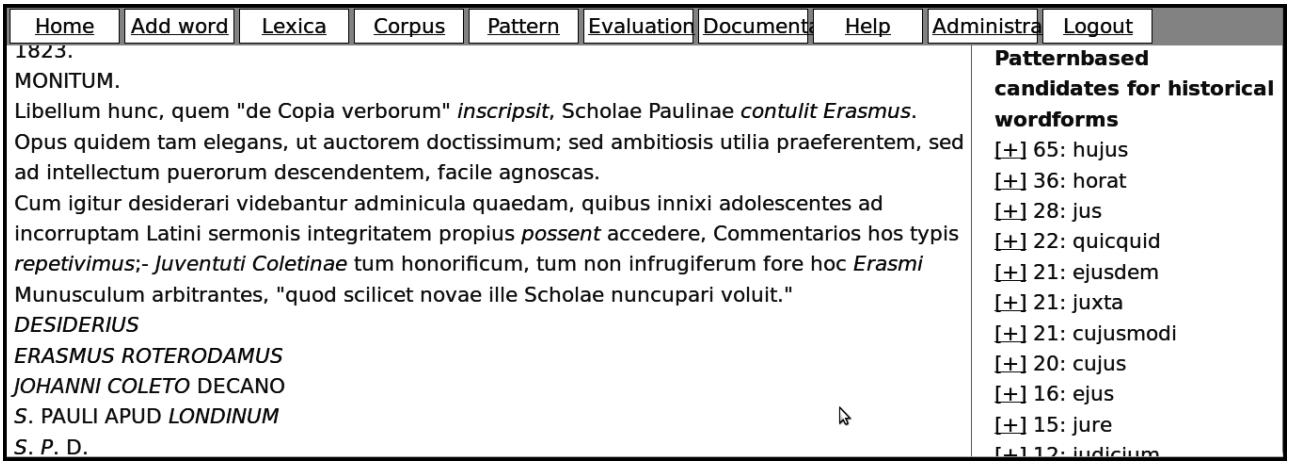


Figure 3: The Lextractor tool for Latin with historical spelling patterns (mainly i → j) indicated in the right column.

are needed to filter out OCR errors. The word coordinates needed for the profiler to work are currently given by ABBYY and Tesseract.

Our immediate next steps consist of getting the lemma lexicon on a firm basis with homogeneous quality regarding its sources and to build out and test the morphology. The lemma lexicon, the lexicon of word forms and the morphology will then be made available under an open source licence. A third lexicon of historical spelling variants will be generated as we get hold of diplomatically transcribed medieval and early modern Latin texts. The promising potential of OCRopus with its new recurrent neural network recognizer will be further explored by training on historical fonts typically employed in early printing and the training file will be made available as well. Finally, having the output of different OCR engines available will enable a voting scheme by which the combined output will be more error-free than each individual result.

8. ACKNOWLEDGMENTS

We thank Markus Brantl of the Bavarian State Library for providing us with high resolution scans of Pexenfelder's book. Thomas Breuel and Adnan Ul-Hasan helped us with OCRopus by generously providing advice and code.

9. REFERENCES

[1] F. Boschetti, M. Romanello, A. Babeu, D. Bamman, and G. Crane. Improving OCR accuracy for classical critical editions. In *Research and Advanced Technology for Digital Libraries*, pages 156–167. Springer, 2009.

[2] T. M. Breuel, A. Ul-Hasan, M. A. Al-Azawi, and F. Shafait. High-performance OCR for printed english and Fraktur using LSTM networks. In *2th International Conference on Document Analysis and Recognition (ICDAR), 2013*, pages 683–687. IEEE, 2013.

[3] M. Federbusch, C. Polzin, and T. Stäcker. *Volltext via OCR - Möglichkeiten und Grenzen: Testszenarien zu den Funeralschriften der Staatsbibliothek zu Berlin - Preußischer Kulturbesitz. Erfahrungsbericht aus dem Projekt "Helmstedter Drucke Online" der Herzog August Bibliothek Wolfenbüttel / von Thomas Stäcker.* Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, Berlin, 2013.

[4] K. E. Georges. *Ausführliches lateinisch-deutsches Handwörterbuch*. Hahnsche Buchhandlung, Hannover u. Leipzig, 1913.

[5] A. Gotscharek, U. Reffle, C. Ringlstetter, and K. U. Schulz. On lexical resources for digitization of historical documents. In *Proceedings of the 9th ACM symposium on Document engineering*, pages 193–200. ACM, 2009.

[6] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[7] J. Leonhardt. *Latein: Geschichte einer Weltsprache*. C. H. Beck, 2009.

[8] C. T. Lewis and C. Short. *A New Latin Dictionary. Founded on the Translation of Freund's Latin German Lexicon Edited By E. A. Andrews, LL D.* Clarendon Press, 1907.

[9] J. Pontanus. *Progymnasmata Latinitatis Sive Dialogi: Vol. 1, De Rebus Literariis*. Sartorius, Ingolstadt, 1589.

[10] J. Rapsch and D. Najock. *Concordantia in Corpus Sallustianum, 2 vols*. Olms, Hildesheim, Zürich, New York, 1991.

[11] S. Reddy and G. Crane. A document recognition system for early modern latin. In *Chicago Colloquium on Digital Humanities and Computer Science: What Do You Do With A Million Books, Chicago, IL*, volume 23. Citeseer, 2006.

[12] U. Reffle and C. Ringlstetter. Unsupervised profiling of OCRed historical documents. *Pattern Recognition*, 46(5):1346 – 1357, 2013.

[13] H. Schmid. A programming language for finite state transducers. In *FSMNLP*, volume 4002, pages 308–309, 2005.

[14] A. Ul-Hasan and T. M. Breuel. Can we build language-independent OCR using LSTM networks? In *Proceedings of the 4th International Workshop on Multilingual OCR*, page 9. ACM, 2013.

[15] T. Vobl, A. Gotscharek, U. Reffle, C. Ringlstetter, and K. U. Schulz. PoCoTo – an open source system for efficient interactive postcorrection of OCRed historical texts. In *DATeCH 2014*. ACM, 2014.