Automatic quality evaluation and (semi-) automatic improvement of mixed models for OCR on historical documents

Uwe Springmann*
Center for Information and Language Processing, LMU; Humboldt-Universität zu Berlin
springmann@cis.uni-muenchen.de

Florian Fink
Center for Information and Language Processing (CIS) LMU
finkf@cis.uni-muenchen.de

Klaus U. Schulz
Center for Information and Language Processing (CIS) LMU
schulz@cis.uni-muenchen.de

ABSTRACT

Good OCR results on historical documents rely on diplomatic transcriptions of printed material as ground truth which is both a scarce resource and time-consuming to generate. A strategy is proposed which starts from a mixed model trained on already available transcriptions from different centuries giving accuracies over 90% on a test set from the same period of time, overcoming the typography barrier of having to train individual models separately for each historical typeface. It is shown that both mean character confidence (as output by the OCR engine OCRopus) and lexicality (a measure of correctness of OCR tokens compared to a lexicon of modern wordforms taking historical spelling patterns into account, which can be calculated for any OCR engine) correlate with true accuracy determined from a comparison of OCR results with ground truth. These measures are then used to guide the training of new individual OCR models either using OCR prediction as pseudo ground truth (fully automatic method) or choosing a minimum set of hand-corrected lines as training material (manual method). Already 40-80 hand-corrected lines lead to OCR results with character error rates of only a few percent. This procedure minimizes the amount of ground truth production and does not depend on the previous construction of a specific typographic model.

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I.7.5 [**Document and Text Processing**]: Optical character recognition (OCR)—*Latin language, historical documents, recurrent neural networks*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Latin language*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

1. INTRODUCTION

In the last years several attempts were made to develop OCR methods for historical documents. As a lot of the early printings have already been digitized (in the sense of making scanned images available), the bottleneck for getting access to the contents of these books now consists in methods of conversion of these images to electronic, machine-actionable text data. Commercial OCR engines lack the possibility to get trained on early typographies and give unsatisfactory results of at most 85% character accuracy on early printings [7, 6, 12, 11], and are deemed completely useless for very early printings (incunabula printings before 1500; [9]).

OCRopus with its new recognizer based upon a recurrent neural network with LSTM architecture has been shown to be trainable on historic fonts and deliver competitive or even better results than either Tesseract or ABBYY on 18th and 19th century Fraktur printings [3]. These results were achieved by generating a lot of training material automatically, generating artificially degraded images from existing text and computer fonts. This method does not work very well for very early printings [11], probably because we lack computer fonts that are similar enough to the actual printings and because the interword spacings are highly irregular, leading to OCR tokens being merged to a single long string without any intervening spaces. Training on real images solves both of these problems [10] but requires the time consuming task of diplomatic transcription of the training material. While the resulting OCR model works very well for the book it has been trained on (often reaching accuracies of 98% for even the earliest printings), these individual models do not generalize well to other books. In an automatic setting where a large amount of books need to be OCRed in short time, the training of individual models is out of the question. The construction of mixed models trained on material from several different books partly overcomes this problem with accuracies still over 90% for a wide variety of books [10], but in the absence of ground truth to test against one cannot know just how good the mixed model is for a particular book or how one could decide whether one model is better than another one without manually counting the errors. An optimal strategy seems to be to start from a mixed model and to refine it later (either automatically or with minimal manual effort), which requires a measure for accuracy independent of ground truth.

In this paper we give (incomplete and preliminary) answers to the following questions:

1. In the absence of ground truth, how can the recognition quality of OCRopus (or any other OCR engine) be estimated automatically?

2. Given an automatic method for OCR quality estimation, how could one use it to construct a model better than the start model *in a fully automated way*, and which improvement can be expected?

3. Which OCR quality can be obtained by adding a small amount of manual work, preparing some lines of ground truth? What is the tradeoff between the number of training lines and the model improvement? How good can it get compared to an individual model trained on a large set of ground truth data from the same document?

Question 1 is of great importance for any large digitization program, where one wants to get a quick estimate of the OCR quality of a book, a page, or a paragraph. To summarize our results:

1. Both *lexicality L*, (a "profiler-based" measure explained below) as well as *mean character confidence C* calculated from individual character confidences of OCRopus (see Fig. 1) correlate with character accuracy. Measurement of either quantity therefore leads to an estimate of the true accuracy. While we find a tighter correlation of accuracy with confidence, lexicality can be calculated for any OCR engine, regardless of the quality of confidence values in the output.

2. For books with an OCR accuracy worse than about 90%, using the prediction of a mixed model as a starting point, appropriate forms of fully automatic training result in an improved quality of about 94% accuracy, confirming the result of [13]). For model selection after training, the above techniques for estimating accuracy are essential.

3. The manual correction of as few as 40-80 text lines and subsequent training on just these lines often leads to excellent models. Even good starting models can be considerably improved. As to the selection of lines, a mixture of randomly selected lines with a set of poorly recognized lines seems to have the best effect. In general, adding more training material leads to better results with diminishing returns, approaching the accuracy of an individual model trained on large amounts of ground truth.

The paper is organized as follows: Section 2 gives a short account of the state of the art for OCR of historical documents to put our work in perspective. Section 3 describes the data sets for our models and experiments. Section 4 describes lexicality and mean character confidence and shows their correlation with accuracy. The last two sections report our experiments and their outcomes for the automatic (Sect. 5) and semi-automatic (Sect. 6) method.

2. RELATED WORK

Work by other groups has mostly focused on Tesseract which is trainable on artificial images generated from computer fonts in a similar way as OCRopus. Training on real data, however, has proved to be difficult, and lead to efforts to reconstruct the original typeset from cut-out glyphs. This has been done by both the Poznań group [4] with their cutouts application (proprietary) and EMOP's Franken+ tool (open source). However, the latter group has reported on reaching only about 86% accuracy on the ECCO document collecction and 68% on the EEBO collection[1]. Their OCR suffers badly from scans of binarized microfilm images containing a lot of noise. A publication from this project with a title similar to the present one [5] consequently deals with improving OCR quality by automatically distinguisting between text and non-text areas. The published "combined models" from this project covering a variety of typesets do not presently give high accuracies.

The Kallimachos project[2] at Würzburg University did have success with Franken+ to reach accuracies over 95% for an incunable printing[3] but this method relies again on creating diplomatic transcriptions from scratch for each individual typeface. The method proposed by [13] to circumvent ground truth production by first training tesseract on a historically reconstructed typeface with subsequent OCRopus training on Tesseract's recognition on the actual book as approximate ground truth has also achieved accuracies above 95% but shifts the effort to the manual (re)construction of the typeface.

A completely different approach was taken with the new Ocular OCR engine by Berg-Kirkpatrick et al. [1, 2] which is able to convert printed to electronic text in a completely unsupervised manner (i.e., no ground truth needed) employing a language, typesetting, inking and noise model. This may be a viable alternative for training individual models with low manual effort, but it seems to be very resource-intensive and slow (transcribing 30 lines of text in 2.4 min according to [2]). Its results are better than (untrained) Tesseract and ABBYY, but it remains to be shown that they consistently reach accuracies higher than 90%.

In summary, while there are other approaches to train individual OCR models for the recognition of historical documents, none have so far reported results as good as OCRopus (consistently over 95% accuracy), nor has it been shown that one could construct generalized models applicable to a variety of books with reasonable results (above 90% accuracy).

3. DATA SETS FOR TRAINING AND EVALUATION - MIXED STANDARD MODEL

The data sets used for training and testing our individual and mixed models consist of Latin books printed with Antiqua types from the 15th to 17th century. We deliberately chose these early printing, among them four incunabula printings from the period 1450 to 1500, because no other OCR methods have been able to yield character accuracies consistently over 95% for such material (see Sect. 1). Scans for these books have been downloaded from archive.org[4] and the Bavarian State Library[5]. The training and testing data consist of individual printed line images extracted from book

[1] `http://emop.tamu.edu/final-report`
[2] `kallimachos.de`
[3] Felix Kirchner, Marco Dittrich, Philipp Beckenbauer, Maximilian Nöth: „OCR bei Inkunabeln – Offizinspezifischer Ansatz der UB Würzburg" (will appear in ABI Technik, Heft 3, 2016)
[4] `http://www.archive.org`
[5] `http://www.digitale-sammlungen.de/index.html?&l=en`

Table 1: Data sets; % error of model prediction

Year		(Short) Title	Author	# lines	label	ind. model	mixed model
1476	∗	Speculum Naturale	Beauvais	2012	1476-S		3.65
1497	∗	Stultifera Navis	Brant (transl. Locher)	1092	1497-S		4.37
1543	∗	De Bello Alexandrino	Caesar	832	1543-D		1.04
1553	∗	Carmina	Pigna	298	1553-C		6.10
1557	∗	Methodus	Clenardus	350	1557-M		10.96
1686	∗	Lexicon Atriale	Comenius	1105	1686-L		5.75
1471		Orthographia	Tortellius	417	1471-O	5.54	10.91
1483		Decades	Biondo	915	1483-D	0.98	11.85
1522		De Septem Secundadeis	Trithemius	201	1522-D	1.53	6.93
1564		Thucydides	Valla	1948	1564-T	1.61	4.32
1591		Progymnasmata vol. I	Pontanus	710	1591-P	3.40	8.81
1668		Leviathan	Hobbes	1078	1668-L	1.89	4.04

pages together with their diplomatic transcriptions serving as ground truth for model training and evaluation of the prediction error rate. These data have been manually compiled over the course of the last two years by one of the authors (US) with some help by students and colleagues. Table 1 gives bibliographic details on these books as well as the amount of data (lines) available. Titles with an asterisk have been used for the training of a mixed model (see Sect. 1).

For training a mixed standard model, about 20% of the available lines for each book have been set aside to form the test set, and the remainder was used as training material (the division was done on a pagewise basis). The resulting model was saved every 1,000 learning steps (each step consists in seeing one line image and its associated ground truth). After training for some thousand steps, the model with the best accuracy on the test set was chosen and its application to the test set gave the prediction error recorded in Table 1. Although our ground truth also contains Greek words which frequently appear in Tortellius and Hobbes, we only trained on Latin glyphs as the common character inventory in all of these books. Greek words are therefore not recognized and all of their characters count as errors. The mixed model was trained by pooling the training sets of all books which contributed to it. The prediction errors of the mixed model on this set of books (from 1% to 11%) are in a similar range as for the other books not represented in its training set, which shows that a mixed model generalizes fairly well over a range of typographies (this is not true for individual models which have high error rates when applied to other books). The mixed model can therefore be taken as a starting point for subsequent model improvements.

The difference in error rate (and correspondingly, accuracy) in Table 1 between the mixed model and the individual models may be seen as the improvement potential for our subsequent experiments to train new individual models with no or minimal manual effort.

4. AUTOMATIC QUALITY EVALUATION

Even for unseen historical documents, the mixed OCR model seems to work nicely, with error rates of a few percent (Table 1). However, in a realistic scenario we want to apply the model to a document where we do not have any ground truth. How can we know if the model works well? This is a core problem in large digitization projects where thousands of books are processed, each having specific problems. The question arises of *how to estimate OCR quality in the absence of any ground truth data?*

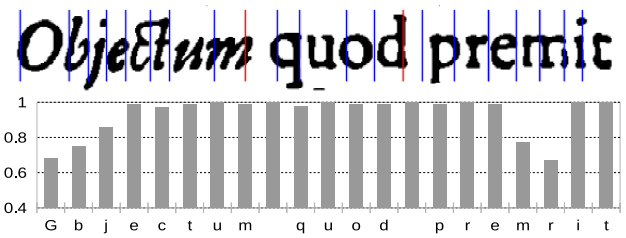
Methods for automatically testing OCR quality are not only important for quality control when we apply a given model. In other scenarios we are faced with a situation where many alternative OCR models are available and we want to find the model which is most appropriate for a given book. This situation is considered below. Furthermore, methods for testing OCR quality can be used to obtain a kind of diagnostics when OCR results for a book do not meet the expectations. Quality testing helps to find those subparts (pages and lines) where serious problems arise, and to find hints on how to improve a model (cf. Figure 3). In a similar way, OCR quality estimates indicate if there is a potential to improve a model, and they can be used to guide the selection of lines to be used as ground truth in model training (s.b.).

We propose two such approximate measures for OCR accuracy: One is the *lexicality* of OCR tokens determined by our language-aware OCR error profiler [8]. For each OCR token the profiler calculates the minimum edit distance (Levenshtein distance) to its most probable modern lexical equivalence, discounting any differences due to historical spelling patterns. The printed word *judicare*, recognized as *judicarc* and with a modern equivalent *iudicare*, gets therefore assigned a Levenshtein distance of 1 (OCR error: e → c), whereas the historical spelling pattern (i → j) is not counted. The sum of these Levenshtein distances over the tokens of a line is therefore a (statistical) measure for the OCR errors of this line, and the lexicality defined as L = (1 - mean Levenshtein distance per character) is a measure for accuracy. Problems with this measure arise from lexical gaps (mostly proper names) and very garbled tokens (either too short such as sequences of single letters, or too long because of merged tokens with unrecognized whitespaces) which do not get Levenshtein distances assigned.

The other measure are the confidence values that OCRopus assigns to its output characters[6]. Whenever an error occurs because one letter gets confused with another similar-looking one, both of them compete for the confidence score and consequently the value assigned to the resulting letter is lower than the values for well-recognized letters. Fig. 1 shows an example: The two lowest confidence values are actual errors (O → G and the insertion of r), other low values correspond

[6]The code of OCRopus had to be slightly adapted to output the confidence value of each character.

Figure 1: Confidence values for characters



to an imperfect recognition (italic b and h look very similar, m is partly recognized as r). More importantly, all characters with a confidence above average (0.93) are correct. The sum of the confidences over all output characters of a line should therefore correlate with the accuracy of this line. Systematic problems for this measure arise from deletion errors (e.g., missed blanks between tokens), because deletions by their very definition do not have a confidence value attached to them.

Below we identify the best model among all the models saved during a training history (every 1,000 learning steps) by choosing the one with the best score (confidence or lexicality). This will work as long as there is a good correlation between these scores and accuracy. The exact relation may be different for each document and even for different training methods on the same document. As soon as some ground truth is available for testing, we can compare the different methods for their actual accuracy. Also, from the statistical properties of the correlation one can give prediction intervals at level α for the accuracy of the OCR result of a complete document based on its measured score x_0 according to the formula:

$$\hat{y} \pm t_{\alpha/2,n-2} S_y \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n \sum x_i^2 - (\sum x_i)^2}} \qquad (1)$$

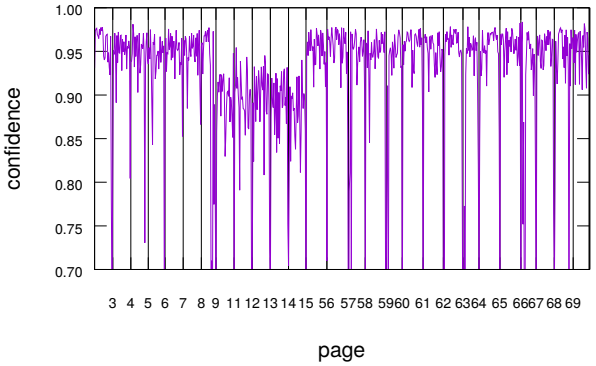
Here, the $(1-\alpha)$-percentile with $(n-2)$ degrees of freedom of Student's T distribution is given by $t_{\alpha/2,n-2}$, and the residual standard error

$$S_y = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} \qquad (2)$$

is calulated from the distances of the y_i from the regression equation $\hat{y} = f(x) = mx + b$ with n data points (x_i, y_i). Fig. 2 shows as an example the correlation of both confidence and lexicality for 1483-D trained on 42 printed lines with ground truth together with their 95% prediction intervals (red lines). Each point corresponds to one model of the training history. We can therefore say that a mean character confidence of 98% for the OCR result leads to an accuracy interval of (95.26%, 95.69%) with 95% probability.

Figure 3 shows another possible application of confidence values for visualizing difficult lines of a document. The document has line headers and footers where the OCR had serious difficulties. For each line we show the average confidence. It is simple to see that lines with low confidence exactly occur at each page break. The reduced confidences in pages 9-15 arise from lines printed in italics which were underrepresented in the trained model.

Figure 3: Visual inspection of OCR quality by a line/page based view of confidence.



5. FULLY AUTOMATIC METHODS FOR IMPROVING OCR ON A GIVEN DOCUMENT

Our fully automatic procedure for improving OCR on a specific document uses two steps.

1. Automatic selection of pseudo ground truth using standardized mixed model. Starting with our standard mixed OCR model (cf. Secction 3) we recognized the given document. Two automatic methods are used to define a collection of "pseudo ground truth" (PGT) lines for training. The first method simply takes the full OCR output as a PGT set. The second method is more complex. Using the information provided by the profiler for each token of the OCR output of the initial mixed model, we looked at tokens where the profiler suggests a correction of certain symbols and at the same time the OCR has low confidence for these symbols. We took all lines containing such a token and replaced the original OCR result by the correction suggestion of the profiler.

2. Automatic training, model evaluation, and new model selection. Using the two types of PGT we started two training runs with OCRopus on the given document. In each run, new OCR-models are saved by OCRopus every 1,000 learning steps. As a result of this automatic training process(es), several alternative OCR-models from two runs are at our disposal, the start model representing one option. Using the average confidence value of all symbols in the OCR output for each model for the given document as a score we chose the OCR model with the best score. The process is illustrated in Figure 4. The x axis gives the sequence number of the models generated during training. Confidence (green), lexicality (blue), and accuracy (purple) show a clear correlation.

Remarks. Note that for the second step any method for automatic quality evaluation could be used. E.g., profiler info (lexicality) might help when using an OCR that does not offer good confidence values. The two steps can be considered as the first round of an Expectation Maximization (EM) algorithm - the OCR output of the mixed model represents a first kind of expected result, and the training is a first optimization step. We could of course iterate both steps, but we did not follow this direction here.

Evaluation of the model obtained. To test the quality of the selected model we used the real GT data. Recall that for evaluation purposes full GT for a significant part of the documents was at our disposal. Using the selected OCR

Figure 2: Predictors for accuracy

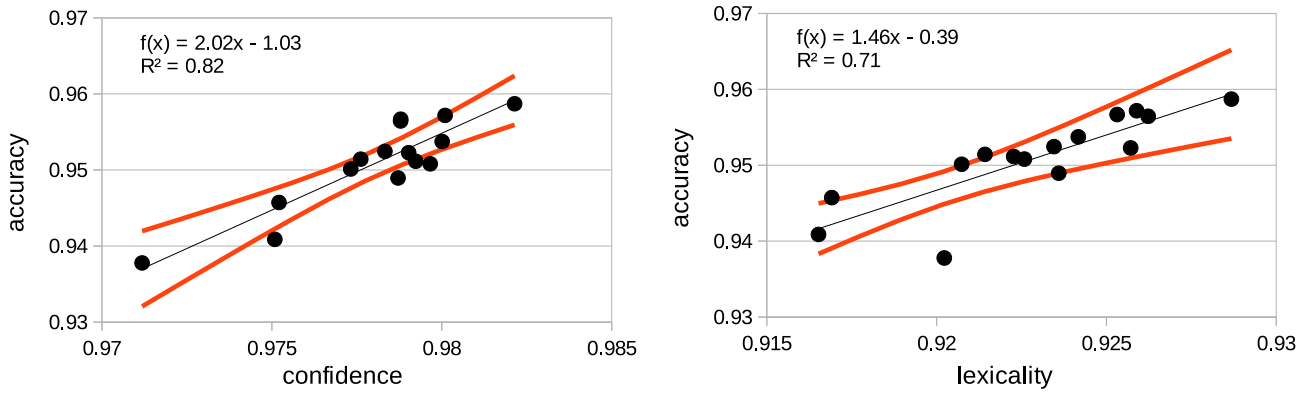
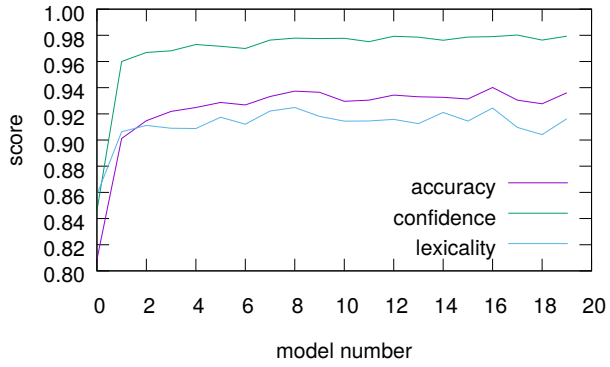


Figure 4: Accuracy, confidence, and lexicality of models generated in a training run.



model we process the document and evaluate the accuracy on this part of the document.

In our test we looked at the three documents from the years 1483, 1522, and 1591 shown in Table 1. In Figure 5 we show the accuracy values reached with the initial standard mixed model (red), with the best model found by exhaustive manual training (black), and with the automatic method (blue). For document 1483, where the initial mixed standard model has a character confusion rate of 11.86%, a remarkable improvement is obtained. For document 1522, where the start model has a high accuracy of 93.07%, no improvement is reached. For document 1591 (start model 91.19%) improvements are modest.

We also checked if better results could be obtained (in theory) when selecting the optimal model from the training processes (as opposed to the model with the best confidence score). Differences are minor.

6. NEARLY AUTOMATIC METHODS FOR IMPROVING OCR ON A GIVEN DOCUMENT

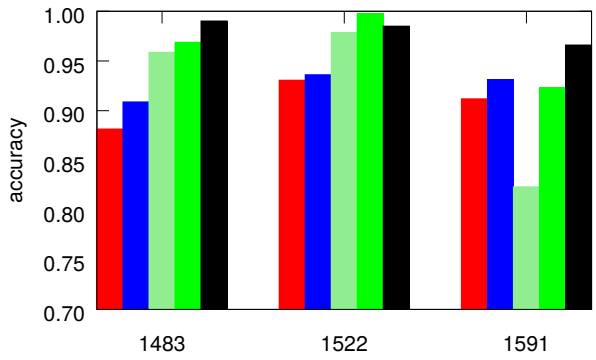
The results above indicate that when starting with a "good" (> 90% accuracy) OCR model for a given document, a fully automatic improvement is often difficult. In order to achieve "excellent" (> 95% accuracy) OCR results for the document a certain amount of manual work seems to be inevitable. We now ask how to obtain a maximal benefit from a minimal amount of manual work. In our case the manual work only consists in the the simple task to transcribe a small number of text lines of the given input document as a GT set for OCR training. Technically, the GT for the lines can be prepared using the OCR output for the mixed model and postcorrecting the selected lines using a postcorrection system [14].

To optimize the benefit (OCR improvement) obtained from transcribing a fixed number of lines and to minimize manual work we looked at different strategies for line selection: (i) selection of a set of consecutive lines, (ii) selection of a random set of lines from the full document, (iii) selection of a collection of lines with high OCR confidence, (iv) selection of lines with low OCR confidence, and (v) mixtures of these strategies. As a first result worth to be mentioned we found that optimal results are obtained when using a mixture of randomly selected lines plus lines with low OCR confidence. A possible explanation is the following: First, random selection of lines have the effect that many distinct pages and positions are taken into account, which is important to obtain improvements on all parts. Second, assuming the lines with low confidence often have many errors, preparing GT for lines with low confidence optimizes the number of positions where model training leads to a real improvement in OCR recognition.

After the selection of the GT material for training, the other steps (training, automatic model selection and evaluation of obtained model) are as above, with the exception that for evaluation of accuracy only lines from a test set have been used that were not part of the previously selected lines for training. Note that it would not be misleading to take the full set of GT lines for computing accuracies since in a real application scenario the transcribed lines are a part of the given document where we want to optimize OCR.

In our tests we again looked at the three documents from the years 1483, 1522, and 1591 shown in Table 1. In two series of experiments, for each document, using the line selection strategy described above we automatically selected (a) 42 lines, (b) 85 lines and trained OCRopus with the GT for this selected set of lines. As a starting point we always used the standard mixed OCR model described above. In Fig. 5 accuracy values reached with 42 lines and 85 lines of manually prepared GT are shown in olive and green, respectively. For document 1483 we come close to the optimized special model

Figure 5: Accuracies reached via automatic (blue) and semi-automatic (green) OCR improvement on documents 1483, 1522, and 1591.



based on a large amount of GT (black). For document 1522, with 85 lines of GT we obtain an almost perfect model that is even better than the optimized special model. For document 1591 the selection of 85 lines leaves room for further optimization.

7. CONCLUSION

In this paper we looked at strategies that help to obtain optimal OCR results on historical documents with a minimal amount of manual work. Summing up, we suggest to use a set of standard mixed models for OCRopus, each covering a spectrum of periods and printings, as a starting point. Standard models could be prepared and exchanged by the community. Once we have such a set, to process a new book we may use an automatic quality measure such as confidence or lexicality (s.a.) to determine the model that offers the best starting point. We may then improve the start model for the given document either in a fully automatic way or by preparing ground truth for a small number of lines.

For finding the best model we again use automatic quality estimation. The results in this paper show that in this way really excellent results can be achieved with a minimal amount of manual work. As a matter of fact, more data and experiments are needed to make this picture more complete and safe. A second important point for future work is to investigate the correlation between confidence or lexicality and accuracy across distinct documents and OCR models.

Acknowledgements. We wish to thank our students Haide Friedrich-Salgado and Jasmin Chebib for creating the ground truth for Hobbes. The material on Tortellius was provided by our Italian colleagues Federico Boschetti, Paola Tomé and Edoardo Bighin, and the ground truth for Brant has been generously shared by the Kallimachos Project at Universität Würzburg (Hans-Günter Schmidt, Felix Kirchner, and Marco Dittrich).

8. REFERENCES

[1] T. Berg-Kirkpatrick, G. Durrett, and D. Klein. Unsupervised transcription of historical documents.

[2] T. Berg-Kirkpatrick and D. Klein. Improved typesetting models for historical OCR. In *ACL (2)*, pages 118–123, 2014.

[3] 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.

[4] A. Dudczak, A. Nowak, and T. Parkoła. Creation of custom recognition profiles for historical documents. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pages 143–146. ACM, 2014.

[5] A. Gupta, R. Gutierrez-Osuna, M. Christy, B. Capitanu, L. Auvil, L. Grumbach, R. Furuta, and L. Mandell. Automatic assessment of OCR quality in historical documents. In *AAAI*, pages 1735–1741, 2015.

[6] M. Piotrowski. *Natural Language Processing for Historical Texts*. Morgan & Claypool Publishers, 2012.

[7] 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.

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

[9] J. A. Rydberg-Cox. Digitizing latin incunabula: Challenges, methods, and possibilities. *Digital Humanities Quarterly*, 3(1), 2009.

[10] U. Springmann and A. Lüdeling. Progress of OCR of early printings exemplified by the RIDGES herbal corpus. *In preparation*, 2016.

[11] U. Springmann, D. Najock, H. Morgenroth, H. Schmid, A. Gotscharek, and F. Fink. OCR of historical printings of Latin texts: problems, prospects, progress. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH '14, pages 57–61, New York, NY, USA, 2014. ACM.

[12] C. Strange, D. McNamara, J. Wodak, and I. Wood. Mining for the meanings of a murder: The impact of OCR quality on the use of digitized historical newspapers. *Digital Humanities Quarterly*, 8(1), 2014.

[13] A. Ul-Hasan, S. S. Bukhari, and A. Dengel. OCRoRACT: A sequence learning OCR system trained on isolated characters. In *DAS2016, to appear*, 2016.

[14] 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 *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH '14, pages 57–61, New York, NY, USA, 2014. ACM.