Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3551–3559
Marseille, 11–16 May 2020

How Much Data Do You Need? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR

Phillip Benjamin Ströbel, Simon Clematide, Martin Volk
University of Zurich
{pstroebel,siclemat,volk}@cl.uzh.ch

Abstract

Recent advances in Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) have led to more accurate text recognition of historical documents. The Digital Humanities heavily profit from these developments, but they still struggle when choosing from the plethora of OCR systems available on the one hand and when defining workflows for their projects on the other hand. In this work, we present our approach to build a ground truth for a historical German-language newspaper published in black letter. We also report how we used it to systematically evaluate the performance of different OCR engines. Additionally, we used this ground truth to make an informed estimate as to how much data is necessary to achieve high-quality OCR results. The outcomes of our experiments show that HTR architectures can successfully recognise black letter text and that a ground truth size of 50 newspaper pages suffices to achieve good OCR accuracy. Moreover, our models perform equally well on data they have not seen during training, which means that additional manual correction for diverging data is superfluous.

Keywords: OCR, neural networks, ground truth, Digital Humanities, black letter, German-language, newspapers

1. Introduction

For non-digital-born textual data, we need Optical Character Recognition (OCR) to extract the text from scanned images. Although OCR systems have improved in recent years, they still result in a certain error rate. It is first and foremost texts in black letter fonts (also called Gothic fonts) which suffer from limited recognition accuracy. The quality of the OCRed material depends on several factors (we provide examples in Figure 1):

- Low distinctiveness of characters, e.g., the "long s" and "f" in older German texts.
- Change over time regarding vocabulary and spelling, e.g., *Commando* (en. *command*) versus *Kommando*, both of which have the same meaning, but come in different forms, thereby enlarging the diversity of possible character sequences.
- Use of small font sizes (and changes in the fonts themselves).
- Mixing of Antiqua and black letter on the same page.
- Distortions of the image due to digitisation from bound newspaper collections or imprecise digitisation workflows.
- Poor paper quality (e.g., smears, smudges, sometimes due to conservation issues, text from the backside shining through, etc.).

Historical documents which have undergone digitisation a few years ago are likely to suffer from low-quality OCR. The possibility of OCR systems to return a confidence score might give an idea about the quality of the extracted text, but such scores are not always *available*, and even if they are, they are often *unreliable*. Additionally, libraries and archives often downscale scanned images when distributing them over the web. For the Digital Humanities, this means dealing with low- to medium-resolution images. In case a researcher wishes to re-run OCR in order to produce a better textual basis, she will be forced to work with this inferior material.

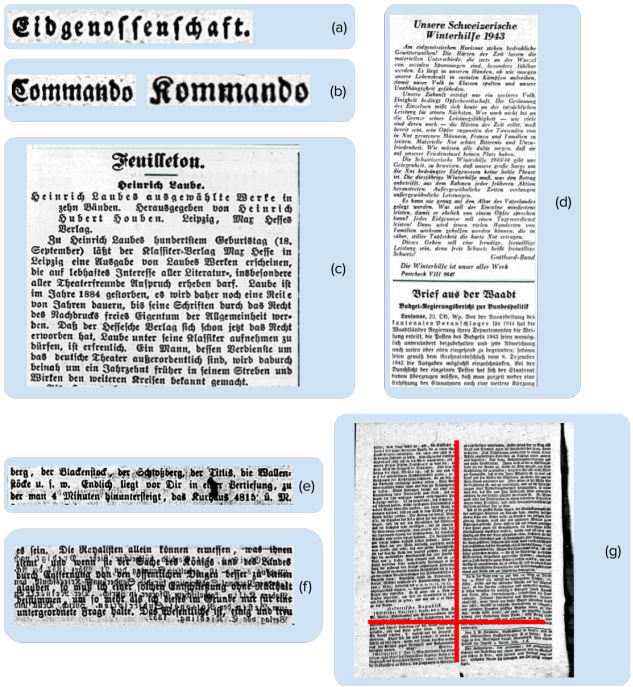

Figure 1: Different factors influencing OCR quality for newspapers: (a) distinctiveness of characters, e.g., an OCR system outputs **Eidgenoffenschast*, **Eidgenossenfchast*, etc., for the German word *Eidgenossenschaft* (en. *confederation*) (b) spelling variations of a word (en. *command*), (c) small (and different) font sizes, different fonts, and small margins between lines, (d) black letter and Antiqua on the same page, (e) smears and smudges, (f) text shining through from the backside, (g) distorted image.

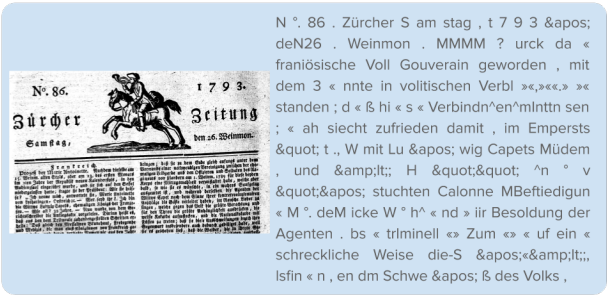

Figure 2: Example OCR from 2005 as produced by the Frauenhofer Institute of a page from 1793.

Having an idea about the quality of the extracted text is not only important for indexing, but also for applying text mining techniques to digitised material. Working with OCRed material means that researchers might base their discoveries on results which contain a certain bias introduced by the OCR (Traub et al., 2015). Often, such errors only become noticeable during the actual work with the digitised corpora. The study by Chiron et al. (2017) took up this point and predicted the risk of missing relevant documents due to OCR errors given a user query in the OCRed Gallica collection. By using the search logs of four months of user queries, they found that 7% of the most common search terms potentially miss relevant documents due to OCR errors. As such, insufficient OCR impedes advances in Digital Scholarship (Smith and Cordell, 2018) due to erroneous text.

Such low-quality OCR is also present in the search portal of the *Neue Zürcher Zeitung*[1] (NZZ), a major German-language newspaper in Switzerland. The Frauenhofer Institute (FI) digitised the stock of the NZZ as early as 2005. Jacob (2005) summarised the process and mentioned that no manual corrections were possible due to monetary restrictions, the wealth of data, and the limited time frame. The FI digitised from microfilm and noticed back then already that the quality of the microfilms differed. This is due to the fact that the NZZ started to microfilm their archives in the 1950s and since then there have been enormous technological advances. The FI tried first and foremost to correct the distortion of images and to enhance unfocused images. These issues were not resolved completely and thus had severe impacts on the quality of the OCR (as Figure 2 shows). Moreover, there was no designated quality control process to evaluate the output of the OCR system.

The need for high-quality OCR is evident and recent advances in machine learning resulted in increasing performance, even on difficult training material. The elaborate deep learning models created new standards in OCR. However, like any machine learning method, deep learning models also need training material. Creating a ground truth (i.e., a set of manually corrected texts) for OCR can be costly and time-consuming. Once a ground truth is ready, the question is which OCR system to use. In this paper, we compare the performance of five different systems. Knowing which system performs best, however, does not directly tell us how much training material is necessary. We, therefore, investigated how the amount of training material affects the performance of OCR. Additionally, we analyse the transferability of OCR models to high-quality images, as well as to other newspapers in order to check whether it is necessary to have several ground truths when we wish to extract the text from different sources.

2. Related Work

The problems accompanying the recognition of text from images of historical documents are manifold and complex, but recent progress in neural OCR techniques has led to significant improvements. Springmann and Lüdeling (2016) analysed the performance of the *OCRopy*[2] system on a diachronic book corpus ranging from 1487 to 1914 (all in black letter fonts). They stated that OCRopy was the only OCR system that achieved character accuracies consistently over 94% and that their models generalised well enough so that they performed equally well on a variety of books. They further claimed that, based on a recommendation by Springmann et al. (2016), a training set of 100 to 200 lines suffices in order to produce comparable results to models which have seen a considerably larger training set.

The *READ* project[3] focuses on the recognition of handwriting. Its core is the *Transkribus*[4] framework. Transkribus is a tool that helps researchers to transcribe manuscripts in the first place. It performs layout analysis which dissects a page into text regions, lines, baselines, and words. Moreover, it offers the possibility to run the *ABBYY FineReader Server 11*[5] OCR software on images uploaded to Transkribus. With the transcriptions made within Transkribus, it is possible to train *Handwritten Text Recognition* (HTR) models (Weidemann et al., 2018). These models are elaborate neural network architectures, which manage to recognise handwriting with only a limited amount of training data. However, we found that HTR models can also be applied successfully for the recognition of printed documents, with character error rates (CER) of 1.1% or better (Ströbel and Clematide, 2019).

As concerns the amount of training data, Martínek et al. (2019) showed that for German-language black letter newspapers, a ground truth of 10 pages is sufficient to achieve a CER of 1.4%. They used a convolutional and recurrent neural network (in a similar, but much simpler fashion than Weidemann et al. (2018)) and different training strategies like, e.g., binarisation of the images, padding, and the use of synthetically generated training data for black letter recognition.

Wick et al. (2018) investigated both the performance of different systems (OCRopy and *Calamari*[6]) on medieval printed books in German and Latin, as well as the amount

[1] https://zeitungsarchiv.nzz.ch

[2] https://github.com/tmbdev/ocropy

[3] https://read.transkribus.eu/

[4] https://transkribus.eu/Transkribus/

[5] https://www.abbyy.com/de-de/finereader-server/

[6] https://github.com/Calamari-OCR/calamari

of training data needed. They found that with increasing number of lines in the training set (starting from only 60) the CER for both systems drastically improved. For example, for data from 1488, their Calamari system managed to lower the CER from 4.9% when using 60 lines to train to 0.43% when using 3,000 lines. However, they do not indicate an optimal number of lines for a training set.

These recent studies show that researchers in Digital Humanities are using a number of different OCR systems to extract text from historical documents. Many systems are freely available. However, it demands a lot of effort from the researchers to understand the different processing steps (Klein et al., 2016), since there is no standardised workflow. *OCR-D* (Neudecker et al., 2019) attempts to facilitate the training of OCR models for collections via creating standards for training and testing a variety of OCR systems instead of limiting users to only one system. Comparisons between different systems are thus simplified and more transparent.

3. Ground Truth Creation

The NZZ was published in black letter from its beginnings in 1780 until 1947[7]. We selected one title page per year from this period at random for manual transcription. By restricting ourselves to title pages we made sure that the main content on the page is textual material. This heuristic guaranteed the exclusion of pages loaded with advertisements of all sorts, stock reports, etc. However, it should be noted that the amount of text on the title page from 1780 (3,180 characters) is much smaller than on a title page from 1939 (23,316 characters), as Figure 3 shows. The more recent data in the ground truth is thus overrepresented, a fact we will take up again in Section 5.2.2.

Figure 3: The number of characters per year of the ground truth from 1780 until 1947.

We extracted the images from the PDFs[8] with *PDFlib TET*[9] and loaded them into Transkribus. We applied the Transkribus internal ABBYY FineReader Server 11 to the collection and used the output text as a basis from which we started the manual correction. We also corrected regions and added baselines, which are required in order to train an HTR model. The resulting ground truth of 167 pages contains 304,286 words and 43,151 lines. Depending on the amount of text on a page, the manual correction of a page takes between 1.5 and 3 hours. We published the ground truth on Zenodo and GitHub (Ströbel and Clematide, 2019).

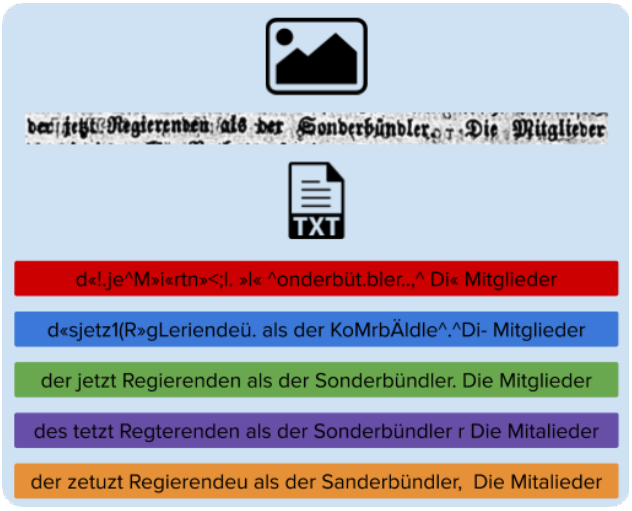

Figure 4: Example outputs from five OCR systems for a line from the NZZ from 1850 (red: FRXIX, blue: FRS11, green: HTR+, purple: kraken, orange: Tesseract).

4. OCR Systems

In our experiments, we evaluate five OCR systems:

1. *ABBYY FineReader XIX*[10] (FRXIX) results from the Frauenhofer Institute provided in 2005,
2. *ABBYY FineReader Server 11* (FRS11) results, available from within Transkribus,
3. a Transkribus HTR+ model (Weidemann et al., 2018),
4. *kraken*[11], and
5. *Tesseract*[12].

Figure 4 shows an example output from these five OCR systems. ABBYY's OCR systems are commercial. We extend the category of non-commercial OCR software with HTR+, kraken, and Tesseract, each of which we present here briefly.

HTR+ models are part of the Transkribus framework, where users have the possibility to train such models given a ground truth. The model proposed by Weidemann et al. (2018) is a deep neural network which combines Bidirectional Long-Short Term Memory (Schuster and

[7]We find newspaper pages in Antiqua before 1947, and some black letter text has been published even after 1947.

[8]PDF was the output format in which the result of the digitisation process from 2005 was saved.

[9]https://www.pdflib.com/de/produkte/tet/

[10]https://www.frakturschrift.com/de:products:finereaderxix

[11]http://kraken.re/

[12]https://github.com/tesseract-ocr/

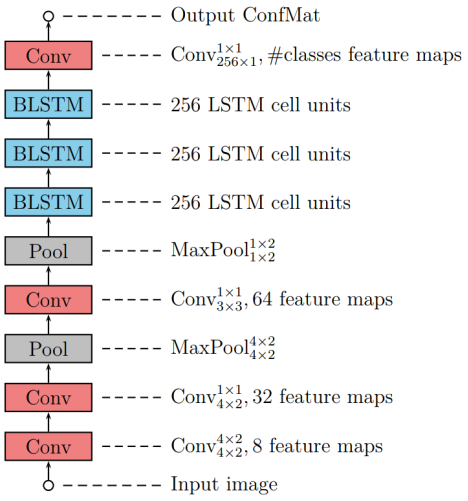

Figure 5: The HTR+ model architecture as presented in Weidemann et al. (2018).

Paliwal, 1997) (BLSTM) layers with convolutional (Krizhevsky et al., 2012) layers. Figure 5 (taken from Weidemann et al. (2018)) summarises the HTR+ model. We used Transkribus version 1.8.0.

kraken is a forked version from OCRopy and supports the training of OCR models with neural networks. In contrast to an HTR+ model, kraken allows the user to determine the architecture of the neural network with the help of the Variable-size Graph Specification Language[13] (VGSL). This means a user has more control over the structure which allows for tuning a model until it delivers good enough results. kraken runs on GPUs, which speeds up training time. We used version 2.0.5.

Tesseract is Google's solution of a neural OCR system. Like kraken, it allows the user to specify the structure of the neural network with VGSL. In contrast to kraken, Tesseract is not GPU-enabled. We used version 4.00.

5. Experiments and Results

In our experiments, we examine the OCR quality of the black letter era of the NZZ. We investigate to what extent different OCR systems improve the OCR quality. Firstly, we are interested in the quality of the OCR provided by the NZZ (produced in 2005). Secondly, we want to scrutinise the amount of training data needed for good OCR quality. Thirdly, the last experiment focuses on the impact of high-quality images on the OCR results in order to judge the transferability of the OCR models.

5.1. Evaluation

We use the bag-of-words F1-measure metric of PRImA TextEval 1.4 (Clausner et al., 2016) for the evaluation the OCR systems' performances. By applying a bag-of-words approach, possible differences in layout recognition or word order of the different systems cannot distort the results.

[13] https://github.com/mldbai/tensorflow-models/blob/master/street/g3doc/vgslspecs.md

If we consider the example in Figure 6, we first determine the bag-of-words for each page in the ground truth as well as for the output of the OCR system, which is kraken in this case. We build the bag of unique words from the respective bags-of-words. This bag determines the number of words the OCR system identified correctly, i.e., true positives (TP). We can now compute Recall, which is the percentage of words the system should have identified, as true positives divided by the sum of true positives and false negatives (FN):

$$R = \frac{TP}{TP + FN} \rightarrow \frac{4}{8} = 0.5. \quad (1)$$

Precision, on the other hand, gives us the percentage of correctly identified words, which is the difference between true positives and the sum of true positives and false positives (FP):

$$P = \frac{TP}{TP + FP} \rightarrow \frac{4}{9} = 0.44. \quad (2)$$

The F1-measure is the harmonic mean between Precision and Recall, as given in Equation 3:

$$F1 = \frac{2 * R * P}{R + P} \rightarrow \frac{2 * 0.5 * 0.44}{0.5 + 0.44} = 0.46. \quad (3)$$

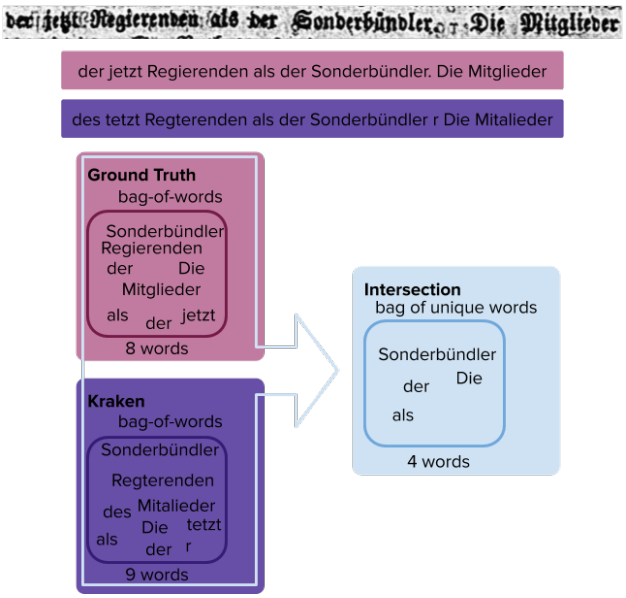

Figure 6: Example for bag-of-words evaluation with one line from which the kraken system extracted the text.

The numbers reported in this paper correspond to the average F1 score of all pages in the test set.

5.2. Experiment 1 — Comparison

The first experiment compares the performance of five different OCR outputs in order to get a quality estimate on the NZZ's original OCR and to investigate how well state-of-the-art methods perform on NZZ material.

Figure 7: Line image-text pair used as training data.

5.2.1. Setup

We extracted the text of the original OCR dating back to 2005 from the PDFs provided by the NZZ. To test FRS11 we applied OCR to the NZZ from within Transkribus. For the models which we train ourselves, i.e., HTR+, kraken, and Tesseract, we used a 90/10[14] split for training and testing, i.e., 150 pages for training, and 17 pages for testing. The test set contains a page from roughly every decade in order to get an apt representation of the diachronicity of the ground truth, and we use this test set throughout our experiments to evaluate the performances of the different systems. The training set of 150 pages comprises a total of 273,440 lines which translates to 38,756 lines. We use the medium-resolution images extracted from the PDFs produced in 2005. These images stem from scanned microfilms of the NZZ. Tesseract and kraken needed more preprocessing. We used the layout information of line regions from FRS11 generated within Transkribus and extracted corresponding line image-text pairs (see Figure 7). These pairs served as input data for training kraken as well as Tesseract.

The set-up of the training process is much simpler for kraken than for Tesseract, since Tesseract expects the transformation of the data into an adequate input format. kraken, on the other hand, works with corresponding .png and .txt files directly.

The HTR+ model trained for 200 epochs with a batch size of 16 and Dropout of 0.5 between the layers. It normalises the input height of the images to 64 pixels. Weidemann et al. (2018) stated that deep neural networks usually need a lot of data to generalise well. Since ground truth creation for handwritten documents is a labourious process, they employed data augmentation to synthetically enlarge the training set. These transformations comprise translation, scaling, rotation, shear mapping and compositions of these methods. We did not apply such methods to the input data for Tesseract and kraken.

For Tesseract and kraken we used the standard settings. The network specifications were:

```
Tesseract = [1,36,0,1 Ct3,3,16 Mp3,3 Lfys48 Lfx96 Lrx96
Lfx256 O1c]
kraken = [1,48,0,1 Cr3,3,32 Do0.1,2 Mp2,2 Cr3,3,64
Do0.1,2 Mp2,2 S1(1x12)1,3 Lbx100 Do 01c]
```

In their essence, both systems work in a similar way. Differences lie in the normalisation of the input (48 pixels for kraken and 36 pixels for Tesseract), the addition of Dropout `Do` for regularisation purposes in kraken, and the use of bidirectional layers `Lb` in kraken. Tesseract, on the other hand, uses two forward LSTMs `Lf` and two reverse LSTMs `Lr` as default setting. `Mp` stands for max pooling and `S` reshapes the output of the previous layer. As concerns hyper-parameters settings Tesseract uses a default learning rate of 0.002, while kraken uses 0.001. The default optimisers for both systems is Adam.

Additionally, we tried to imitate the HTR+ model from Figure 5 in VGSL in kraken (we call the rebuild kraken+) in the following way:

```
kraken+ = [256,64,0,1 Cr4,2,8,4,2 Cr4,2,32,1,1 Mp4,2,4,2
Cr3,3,64,1,1 Mp1,2,1,2 S1(1x0)1,3 Lbx256 Do0.5 Lbx256
Do0.5 Lbx256 Do0.5 Cr255,1,85,1,1]
```

We used kraken because of the much easier handling as compared to Tesseract. We applied Dropout after every BLSTM layer to prevent overfitting. The number of filters in the last convolutional layer of the HTR+ depends on the number of different characters. Since we cannot control for this number in VGSL, and since we could not find this information in Weidemann et al. (2018), we examined the frequency distribution of the characters in the training set. We determined that limiting the number of filters to 85 (out of 144 different characters, among which we find also symbols) is a reasonable threshold. We used a batch size of 256 and the same normalisation of the input height. We reduced the learning rate to 0.0001. Moreover, we used a cyclical learning rate (Smith, 2017), which should result in faster convergence of the model and lead to better results than a fixed learning rate or a learning rate schedule without having to optimise for the best learning rate. With the cyclical learning rate, the kraken model did not show any significant improvements on the validation set after 18 epochs.

We evaluated the output of the six different systems with the bag-of-words F1-measure presented in Section 5.1.

5.2.2. Results of Experiment 1

Table 1 summarises the performance of the five OCR systems as the average of all F1 scores over the test set. Unsurprisingly, the results of the 2005 FRXIX system are far worse than those of the new systems. For the page from 1820, the F1-measure of the FRXIX output is at only 38.2% and thus totally illegible. A closer look at the page shows that it indeed poses a problem due to the very bad image quality and large portions of the text from the backside shining through. FRS11, i.e., a pure re-OCRing using ABBYY's most recent version, achieves an improvement of almost 13 percentage points. The result of HTR+, however, leads to an additional improvement of more than 16 percentage points. That the F1 scores for kraken and Tesseract are almost the same was to be expected since they use comparable (although not identical) architectures. Tesseract's results are slightly more consistent though, but kraken's use of a BLSTM layer seems to be beneficial. If we look at the HTR+ rebuilt model using kraken, we find that it performs only 2.24 percentage points short of the HTR+ model. As such, kraken is an important candidate to consider for building an OCR system. Last but not least, and referring back to the imbalanced word distribution over the years mentioned in Section 3, the overrepresented data from later periods of the time span for which we have constructed a ground truth does not pose a problem for the recognition of text from pages belonging to the part of the

[14]Typically, these systems use 10% of the training data as validation set during training.

underrepresented early data.

	FRXIX	FRS11	HTR+	kraken	Tesseract	kraken+
F1	0.678	0.811	**0.978**	0.865	0.861	0.954
SD	0.111	0.073	0.013	0.064	0.055	0.024

Table 1: Average mean and standard deviation of the performance of six OCR systems on the test set.

5.3. Experiment 2 — Ablation

The ablation experiment aims at determining the number of pages needed during training an OCR system to achieve the best results possible. The results from subsection 5.2.2 indicate to focus on HTR+ and kraken+ for the ablation experiment.

5.3.1. Setup

For this experiment, we split the corpus in different training set sizes, while the test set remains the same. Table 2 provides an overview of the different training test sizes. We decided to test training set sizes of 12, 25, 50, 100, and 150 pages, where bigger training sets always contain the pages from smaller training sets. One exception is HTR+ 200L, where we sampled 200 lines at random from training set pages[15]. We chose to train a model on 200 lines, because Springmann and Lüdeling (2016) reported from their experiments that the character accuracy ranges from 95% to sometimes even 99%. We then trained an HTR+ model within Transkribus for each training set size for 200 epochs and again compared the performance of each model on the test set using the bag-of-words F1-measure. We carried out the same ablation with training set sizes from 12 to 150 pages with kraken. From the learning curve of the HTR+ model in Section 5.2.2 (see also Figure 8) we determined 50 epochs as sufficient to train the kraken+ models. While for the biggest training set size the HTR+ model takes about 8 hours, kraken took more than two days to run on the same training set on our infrastructure.

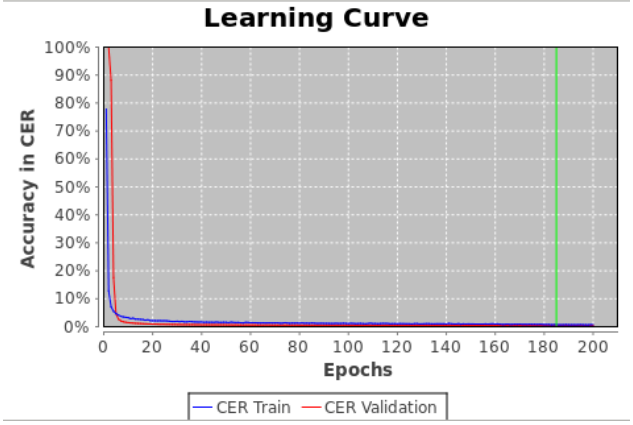

Figure 8: Learning curve over 200 epochs for HTR+ model.

[15]We used the sampler integrated into the Transkribus tool.

Training set size	# words	# lines
150 pages	273,440	38,756
100 pages	183,246	25,904
50 pages	89,705	12,740
25 pages	45,134	6,404
12 pages	22,081	3,157
200 lines	1407	200

Table 2: Summary of all training sets used for the training of different models divided in number of words and number of lines (we used the 200 lines for HTR+ only).

5.3.2. Results of Experiment 2

We first notice from Table 3 that a training set of only 200 lines is not sufficient. We also point to the fact that as soon as character accuracy drops below a certain threshold, word accuracy suffers significantly. As such, we deem the bag-of-words F1-measure to be more telling than character accuracies, but we nevertheless include the character error rate (CER) of the HTR+ models for informative purposes. The HTR+ 200L model, with a CER of 5.04%, achieves an average bag-of-words F1-score of 0.808, which is only slightly worse than FRS11 (see Table 1). In general, we note an upwards trend, i.e., the more training lines we include, the better the models perform, both in terms of CER and bag-of-words F1-measure. Also, the SDs are so low that we can conclude that the models output consistent results over all periods.

	Models					
	HTR+ 150	HTR+ 100	HTR+ 50	HTR+ 25	HTR+ 12	HTR+ 200L
CER in %	**0.48**	0.50	0.56	0.68	1.10	5.04
F1	**0.978**	0.977	0.972	0.965	0.954	0.808
SD	0.013	0.012	0.014	0.017	0.022	0.066

Table 3: Performance of HTR+ using different training set sizes (200 lines, as well as 12, 25, 50, 100, and 150 pages).

Table 4 makes a reference to the number of pages needed for a ground truth in order to produce high-quality OCR, since we were interested in whether there is a significant difference in the performance of the models when trained on differently sized training sets. In order to be able to make a statement about these differences, we performed a One-Way ANOVA on the 17 bag-of-words F1 scores of the test set. We excluded results of the HTR+ 200L model since its scores differed too much from those of the other systems (the results would remain the same). We see a significant performance drop between HTR+ 50 and HTR+ 25 at a significance level of $p = 0.0001$, which is highly significant. We conclude from this result that a training set size of 50 is sufficient for high-quality OCR, and that adding pages for training (e.g., 50 or even 100 more) does not lead to significant performance boosts.

	Models				
	kraken+ 150	kraken+ 100	kraken+ 50	kraken+ 25	kraken+ 12
CER in %	**0.89**	0.99	1.40	1.62	3.61
F1	**0.953**	0.946	0.924	0.905	0.821
SD	0.024	0.029	0.041	0.047	0.072

Table 5: Performance of kraken+ using different training set sizes (12, 25, 50, 100, and 150 pages).

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
HTR+ 150	17	16.6196	0.9776	0.0002
HTR+ 100	17	16.6111	0.9771	0.0001
HTR+ 50	17	16.5307	0.9724	0.0002
HTR+ 25	17	16.4028	0.9649	0.0003
HTR+ 12	17	16.2148	0.9538	0.0005

Table 4: Significance testing in order to determine the optimal number of pages in the training set. The red line indicates where the difference between groups becomes significant.

The results for the kraken+ models show a similar picture. Although the overall scores in terms of bag-of-words F1-measure for the kraken+ models are lower than those for HTR+, Table 5 shows the same trends, namely that the performance of the models improves with increasing training set size. We hypothesise that the data preprocessing and augmentation, i.e., contrast normalisation, skew correction, slant normaliser, affine transformation, dilation, erosion, as well as elastic and grid-like distortions, as mentioned in Weidemann et al. (2018), lead to an enhanced performance.
We applied the same One-Way ANOVA statistics to the kraken+ models and found the same division as for the HTR+ models at a significance level of even $p < 0.0001$, hereby confirming that a training set size of 50 pages produces accurate OCR results. This is convenient, especially if we take into account the training times of the kraken+ models (see Table 6). As such, it is possible to train reliable OCR systems within 13.5 hours. We should also point out that training times for HTR+ models with 8 hours on the biggest training set size is much lower (our kraken+ model took roughly 43 hours to complete training on the biggest training set). Nevertheless, our results also indicate that it is still possible to achieve better results by adding more material, since the F1 scores are higher for models trained with more data. However, the results also tell us that in terms of time and cost optimisation for ground truth annotation, 50 pages are sufficient. We could have saved between 200 and 300 hours of work, if we had known this beforehand.

	Models				
	kraken+ 150	kraken+ 100	kraken+ 50	kraken+ 25	kraken+ 12
Training time	43h 09min	28h 02min	13h 31min	9h 31min	3h 12min

Table 6: Training times for different training set sizes over 50 epochs on GeForce GTX TITAN X.

5.4. Experiment 3 — Transfer

This experiment tests whether the application of the models we have trained on medium-quality NZZ images remains stable when we apply them to high-quality images on the one hand, and to pages printed in black letter from other newspapers on the other hand.

5.4.1. Setup

In order to test the former case, we take high-quality scans digitised directly from paper for four pages from the NZZ test set (1780, 1830, 1880, 1929). We then extracted the text from these pages using the HTR+ models 12 to 150 and also compared these results to OCR results from FRS11. We tested the transfer to other German-language and black letter publications by taking five high-quality scans each from the *Bundesblatt* and the *Neue Zuger Zeitung*[16] covering the last 50 years of the 19th century and extracted the text using the same systems. Again we used the bag-of-words F1-measure to evaluate the performance on the slightly different data.

5.4.2. Results of Experiment 3

First, we looked at the average of each model on a reduced 4-pages test set (all of which are from the original test set). The overall picture (see Table 7) is that the performance stays about the same. For the recognition of text in high-quality images, there are slight deteriorations of 0.03 and 0.01 for the HTR+ 150 model and the HTR+ 50 model, respectively. FRS11 profited the most from the better image quality and improved almost 5 percentage points. The results in Table 7 suggest that the models are transferable and that the performance will not suffer, although we have trained the models on medium-quality images.

As concerns the transferability to other publications, we see average F1 scores of above 95% for all HTR+ models (see Table 8). Moreover, FRS11 scores over 90% for the first time on the Bundesblatt. It is, however, still 5 percentage points away from the best performing HTR+ model for the same data. The results show that a transfer between publications is possible and that the OCR at times is even better than on the data the systems saw during training. We should mention, though, that the standard deviation (SD) is a little higher in these evaluations, which stems from the smaller test set size.

[16]The *Bundesblatt* is a collection of announcements by the Swiss government. The *Neue Zuger Zeitung* is another German-language Swiss newspaper, not to be confused with the NZZ.

	Paper scans		Microfilm	
Models	F1	SD	F1	SD
HTR+ 150	0.982	0.002	**0.985**	0.005
HTR+ 100	**0.983**	0.004	0.983	0.005
HTR+ 50	0.977	0.004	0.978	0.006
HTR+ 25	0.972	0.009	0.969	0.008
HTR+ 12	0.969	0.005	0.956	0.011
FRS11	0.891	0.031	0.847	0.024

Table 7: Mean F1 and standard deviation (SD) of the HTR+ models and FRS11 on four high-quality images (1780, 1830, 1880, 1929) of the NZZ, compared to performance on scans from microfilm.

	Bundesblatt		Neue Zuger Zeitung	
	F1	SD	F1	SD
HTR+ 150	**0.986**	0.015	0.990	0.007
HTR+ 100	**0.986**	0.013	**0.991**	0.006
HTR+ 50	0.982	0.01	0.989	0.007
HTR+ 25	0.982	0.015	0.985	0.009
HTR+ 12	0.976	0.018	0.982	0.010
FRS11	0.924	0.027	0.884	0.037

Table 8: Average performance on five pages each from other publications.

6. Conclusion

It became evident during our work that character error rates (CER) are a good indicator about the models' ability to identifying characters correctly. However, for any further data processing which may include indexing or applying text mining techniques, the bag-of-words F1-measure provides a better picture of the systems' performances.

6.1. Experiment 1

We were able to show that state-of-the-art neural OCR systems like HTR+, kraken, and Tesseract trained on in-domain data perform better than standard commercial systems. The kraken results are on par with Transkribus but kraken provides more freedom when training an OCR model.

With Transkribus, there are still restrictions, e.g., it is not possible to download the model in order to apply it independently from the Transkribus infrastructure. Moreover, although the creation of a ground truth within Transkribus is simple, annotators need to add baselines in order for an HTR+ model to work. This is obsolete for kraken[17] (and Tesseract), which only require line images without baseline information.

The difference between e.g., HTR+ 150 and the kraken+ is likely to originate in Transkribus' usage of baseline information, as well as the various preprocessing and data augmentation steps.

Finally, it is also possible to do ground truth annotation with kraken, as well as image segmentation. Future work should analyse the impact of using kraken's line segmentation tools. First trials did not show the expected quality, however.

6.2. Experiment 2

We furthermore provided an estimate of the amount of training pages necessary to provide solid OCR results using HTR+ and kraken+. Our experiments show that for both systems a ground truth of 50 pages is sufficient to provide good OCR results. This insight significantly reduces the time needed to create a ground truth.

6.3. Experiment 3

We have shown that the models are transferable to other periodicals or newspapers. Further research should scrutinise whether tuning an already existing model (e.g., HTR+ or kraken+) with pages from "unknown" material results in another performance boost.

[17] more recent versions of kraken offer the possibility to train with baseline information, though

7. Acknowledgements

This research has been supported by the Swiss National Science Foundation under grant CR-SII5_173719. We would also like to express our gratitude to Günter Mühlberger and Max Weidemann from the READ project for their support. Moreover, we thank Camille Watter and Isabel Meraner who helped in the ground truth annotation process. We are especially grateful that the NZZ gave us access to their data collection. Last but not least, our thanks go to Benjamin Kissling from kraken.

8. Bibliographical References

Chiron, G., Doucet, A., Coustaty, M., Visani, M., and Moreux, J.-P. (2017). Impact of OCR errors on the use of digital libraries: towards a better access to information. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*, pages 249–252. IEEE Press.

Clausner, C., Pletschacher, S., and Antonacopoulos, A. (2016). Quality prediction system for large-scale digitisation workflows. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 138–143. IEEE.

Jacob, K. (2005). 70 Terabyte Zeitgeschichte. *Frauenhofer Magazin*, 2.

Klein, E., Däumer, M., and Hütig, A. (2016). Gute Ergebnisse aus „schlechten" Textvorlagen. *Information-Wissenschaft & Praxis*, 67(5-6):331–338.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Martínek, J., Lenc, L., and Král, P. (2019). Training strategies for OCR systems for historical documents. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 362–373. Springer.

Neudecker, C., Baierer, K., Federbusch, M., Boenig, M., Würzner, K.-M., Hartmann, V., and Herrmann, E. (2019). OCR-D: An end-to-end open source OCR framework for historical printed documents. In *Proceedings of the 3rd International Conference on*

Digital Access to Textual Cultural Heritage, pages 53–58. ACM.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Smith, D. and Cordell, R. (2018). A research agenda for historical and multilingual optical character recognition. Technical report, Northeastern University.

Smith, L. N. (2017). Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE.

Springmann, U. and Lüdeling, A. (2016). OCR of historical printings with an application to building diachronic corpora: A case study using the ridges herbal corpus. *arXiv preprint arXiv:1608.02153*.

Springmann, U., Fink, F., and Schulz, K. U. (2016). Automatic quality evaluation and (semi-) automatic improvement of OCR models for historical printings. *arXiv preprint arXiv:1606.05157*.

Ströbel, P. B. and Clematide, S. (2019). Improving OCR of Black Letter in Historical Newspapers: The Unreasonable Effectiveness of HTR Models on Low-Resolution Images. In *Proceedings of the Digital Humanities 2019, (DH2019)*. CLARIAH.

Traub, M. C., Van Ossenbruggen, J., and Hardman, L. (2015). Impact analysis of OCR quality on research tasks in digital archives. In *International Conference on Theory and Practice of Digital Libraries*, pages 252–263. Springer.

Weidemann, M., Michael, J., Grüning, T., and Labahn, R. (2018). HTR Engine Based on NNs P2 Building Deep Architectures with TensorFlow. Technical report.

Wick, C., Reul, C., and Puppe, F. (2018). Comparison of OCR accuracy on early printed books using the open source engines calamari and ocropus. *JLCL*, 33(1):79–96.

9. Language Resource References

Ströbel, Phillip Benjamin and Clematide, Simon. (2019). *NZZ Black Letter Ground Truth*. University of Zurich, 1.0, ISLRN 855-418-004-842-0.