mögliches Journal: Language Resources and Evaluation (Springer)
https://www.springer.com/journal/10579/
https://www.springer.com/journal/10579/submission-guidelines
Full-length paper: 18-25 S., Word Doc

Do We Need Clean OCR?

The Impact of OCR Corrections on Named Entity Recognition in German-Language Newspapers

Sina Menzel [0000-0003-1798-2672][1], Hannes Schnaitter [0000-0002-1602-6032][1], Clemens Neudecker [0000-0001-5293-8322][2], Vivien Petras [0000-0002-8113-1509][1] vivien.petras@ibi.hu-berlin.de, Josefine Zinck [0000-0002-7511-098X][1],
Kai Labusch [0000-0002-7275-5483][2]
[1] Berlin School of Library and Information Science, Humboldt-Universität zu Berlin, Germany
[2] Berlin State Library, Prussian Cultural Heritage Foundation, Germany

Abstract. (150-250 words)

Keywords. Optical Character Recognition, Ground Truth, Named Entity Recognition, evaluation, historical newspapers

Availability of data, code and materials.

- NEAT, NER/NEL → github
- annotation guidelines (publish on Zenodo)
- ground truth NER (publish on Github + Zenodo)

Status

1. ~~Introduction Vivien~~
2. SOTA - Hannes
3. Corpus + NER pipeline - Clemens, Kai, Vivien
4. NEAT - Clemens, Kai
5. Entity Types, annotation guidelines, agreement - Vivien
6. Sample Corpus - Hannes
7. Results - Hannes
8. Conclusion - Vivien
9. ~~Acknowledgements~~
10. References

1. Introduction

The amount of machine readable historical sources has significantly increased over the past 20 years due to extensive digitization campaigns of print material by cultural heritage institutions (Kaplan and di Lenardo 2017). At the same time, this material gathers more attention in information retrieval research (Ehrmann et al. 2020a). In order to make digitized text machine-processable, it needs to be recognized as text and mapped from images to computer-readable character codes. With optical character recognition (OCR) in place, text processing prepares the text for improved access, i.e. tokenization, part-of-speech tagging or named entity recognition and linking (NER / EL). However, the OCR process introduces errors in the recognized text, particularly in historical sources (Kugler 2018, p. 42). Accuracy scores on character detection decrease significantly on historical texts caused by difficulties in recognizing gothic lettering, poor initial quality in paper and ink or aging blemishes on the original paper. Nevertheless, machine learning algorithms have improved by such a degree in recent years - particularly with the introduction of deep learning with neural networks - that the quality of the source material appears to have less impact on the success of the text processing task. In this paper, we look at the potential of deep learning algorithms to compensate for erroneous character recognition by testing the impact of different levels of OCR quality on named entity recognition (NER). In particular, we study whether NER improves with corrected OCR texts as compared to uncorrected OCR texts when using the BERT language model on German historical newspaper texts. The research questions guide the analysis from an overall quality assessment to more detailed aspects:

RQ1: How effective is BERT NER on historical newspaper material in German?
RQ2: Does NER quality differ if we correct OCR or not?
RQ3: Does NER quality differ by entity type and OCR correction?

Additional to the quality assessments, this study also contributes other valuable resources to the language technology research community:

- a ground truth corpus of intellectually coded and annotated named entities in German-language historical newspapers,
- guidelines for the detection and annotation of named entities in historical newspaper texts, and
- a software tool (NEAT), which supports OCR corrections and manual annotations of named entities in newspapers.

Error categories and challenges occurring in NER in a historical newspaper corpus will also be discussed.

The paper is organized as follows: section 2 describes the state of the art for NER, especially in historical texts. Section 3 details the newspaper corpus and the OCR approach used. Sections 4 and 5 describe the annotation tool NEAT, which was developed for the project, and the manual annotation process and its guidelines for the development of the ground truth corpus. Finally, section 6 presents the results of the comparative analysis and section 7 concludes with recommendations for further work.

2. OCR and NER for Historical Texts - Hannes

The number of openly available textual datasets in different languages has increased as more and more historical corpora are digitized. NER is one of the primary applications for information retrieval as it identifies relevant tokens in the text. First definitions of entity recognition emerged in the early 1990s and information retrieval research introduced evaluation tasks to track progress in the field as early as 1995 with MUC-6[1], followed by CoNLL[2] 2002 and 2003, as well as CLEF-ER[3] 2013. One significant outcome of these tasks was the strong focus on entities of the three most frequent classes of proper

[1] 6th Message Understanding Conference. https://cs.nyu.edu/faculty/grishman/muc6.html
[2] 4th and 5th Conference on Computational Natural Language Learning. https://www.clips.uantwerpen.be/conll2002/ner/; https://www.clips.uantwerpen.be/conll2003/ner/
[3] 4th Conference and Labs of the Evaluation Forum, http://clef2013.clef-initiative.eu/

names: persons, locations and organizations. This traditional set was challenged by Grouin et al. 2011, who introduced an extended taxonomy, including numerical types such as time, date and amount.

- NER projects
    - Full text searches were one of the central elements of the QUAERO[4] project, a large Franco-German initiative for the development of a search engine running from 2004 to 2013. For the direct OCR-processing of digitized greyscale texts, this resulted in the development of a tool called *Grey Level Character Recognition System*[5]. Furthermore, named entity recognition was performed on historical newspapers in French. For an evaluation campaign in 2012 (Galibert u. a. 2012), a large ground truth corpus of 295 pages, the *Quaero Old Press Corpus*[6], was created.
    - One of the main goals of the Swiss project *impresso - Media Monitoring of the Past*[7] is to "*apply text mining techniques to transform noisy and unstructured textual content into semantically indexed, structured, and linked data; to develop innovative visualization interfaces to enable the seamless exploration of complex and vast amounts of historical data*" (Ehrmann et al. 2020c, p. 959). In the scope of the project, different text corpora are released, inter alia an OCR ground truth consisting of 167 manually corrected pages of the *Neue Zürcher Zeitung*, as well as named entity ground truth for historical newspaper texts in different languages, on the basis of which the CLEF 2020 HIPE task was performed (ibd. p. 962; Ehrmann et al. 2020a).
    - Europeana Newspapers (Neudecker et al. 2014, http://blog.kbresearch.nl/2014/03/03/ner-newspapers/)
- NER projects for German:
  https://www.aclweb.org/anthology/P18-2020/
  https://arxiv.org/abs/1906.07592

- OCR correction studied? (Noch einzuarbeiten: Chiron et al. 2017; Hamdi et al. 2019; Linhares Pontes et al. 2019; Lopresti 2009; Neudecker and Antonacopoulos 2016; Packer et al. 2010; Smith and Cordell 2018; Traub et al. 2015; van Strien et al. 2020)

Guillaume Chiron, Antoine Doucet, Micka¨el Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. Impact of OCR errors on the use of digital libraries. In Proceedings of ACM/IEEE-CS Joint Conference on Digital

Libraries, Toronto, Ontario, Canada, June 2017 (JCDL'17), find: among the 300k tokens matched, 8k are affected by OCR errors. 2k common search terms are causing mismatches, meaning that 7% of the queried
terms potentially miss documents due to OCR errors. IR study based on Gallica queries on historic documents
also claim: Historical newspapers for example, due to their complex layout and their original fonts have been reported to be especially challenging for OCR engines with up to 10% of wrongly detected characters on some documents.
15% of the wrongly OCRed terms are Named Entities

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.: The researchers found that the impact of OCR errors is not substantial for a task that compares two subsets of the corpus. For a different task, the retrieval of a list of the most

[4] http://en.www.quaero.org.systranlinks.net
[5] http://en.www.quaero.org.systranlinks.net/module_technologique/grey-level-character-recognition-system/
[6] https://catalogue.elra.info/en-us/repository/browse/ELRA-W0073/
[7] https://impresso-project.ch/

significant words (in this case, describing moral judgement), however, recall and precision were considered too low.

R. Holley. How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine, 15(3/4), 2009.

S. Tanner, T. Mu~noz, and P. H. Ros. Measuring mass text digitization quality and usefulness. D-Lib Magazine, 15(7/8):1082{9873, 2009.

K. Kettunen, T. Honkela, K. Linden, P. Kauppinen, T. Paakkonen, J. Kervinen, et al. Analyzing and improving the quality of a historical news collection using language technology and statistical machine learning methods. In IFLA World Library and Information Congress Proceedings 80th IFLA General Conference and Assembly, 2014.

B. Alex, C. Grover, E. Klein, and R. Tobin. Digitised historical text: Does it have to be mediOCRe? In J. Jancsary, editor, Proceedings of KONVENS 2012, pages 401{409. OGAI, September 2012. LThist 2012 workshop.: There is ample research into how to reduce the error rates of OCRed text in a post-processing phase. For example, removing common errors, such as the \long s"-to-f confusion or the soft-hyphen splitting of word tokens, has shown to improve Named Entity Recognition. This, however, did not increase the overall quality to a succient extent as it addressed only 12% of the errors in the chosen sample

Most studies on the impact of OCR errors are in the area of ad-hoc IR, where the consensus is that for long texts and noisy OCR errors, retrieval performance remains remarkably good for relatively high error rates [17K. Taghva, J. Borsack, A. Condit, and S. Erva. The e
ects of noisy data on text
retrieval. J. Am. Soc. Inf. Sci., 45(1):50{58, Jan. 1994.]. On short texts, however, the retrieval effectiveness drops significantly [7 W. B. Croft, S. Harding, K. Taghva, and J. Borsack. An evaluation of information retrieval accuracy with simulated OCR output. Technical report, Amherst, MA,
USA, 1993., 13E. Mittendorf and P. Schauble. Information retrieval can cope with many errors.
Inf. Retr., 3(3):189{216, Oct. 2000.]. In contrast, information extraction tools suffer significantly when applied to OCR output with high error rates [16 K. Taghva, R. Beckley, and J. Coombs. The e
ects of OCR error on the extraction
of private information. In H. Bunke and A. Spitz, editors, Document Analysis
Systems VII, volume 3872 of Lecture Notes in Computer Science, pages 348{357.
Springer Berlin Heidelberg, 2006.]. Studies carried out on unreliable. OCR data sets often leave the OCR bias implicit

- Different approaches to the post-correction of OCR have been made.
- OCR quality is typically measured based on word or character accuracy indicated by the Levenshtein-distance (Kugler 2018, p. 46 f.).
- The following types of errors might occur (based on Zumstein/Baierer 2016, p. 74-75):

  Type I: Character errors
  The most frequent and most relevant type of error in the annotation process, where there are mistakes in the recognition of characters, often caused by poor scan quality or gothic lettering, e.g., a common error is ſ → f.

  Type II: Segmentation errors
  These errors are a special type of character error, where spaces between tokens are not recognized correctly. This leads to the incorrect splitting or merging of tokens and is often caused by line breaks in the original text.

  Type III: Word errors
  Word errors are character errors of full words. This frequently occurs in correlation with shifting fonts or if automated post-OCR-normalizations apply. The latter are usually based on wordlists that might disimprove individual tokens.

Type IV: Sectional errors
This type refers to formatting errors regarding the layout or other textual sections, e.g., sentence boundaries.

- Rodriquez et al. (2012) evaluated the performance of different OCR-tools on the resulting NER for the traditional entity types (persons, locations, and organizations). The best performing OCR system enabled an F1 score of 0,61 for subsequent entity recognition (p. 412). Furthermore, they performed tests on manual post-ORC correction on their datasets, which did not significantly improve NER-output (p. 413 f.).
- Kettunen and Ruokolainen (2017) examined the effect of OCR quality on NER results. For this, they compared the performance of four NER software tools on historical newspaper text in Finnish with appr. 73% word correctness after OCR (p. 184).
- Alex and Burns (2014) introduced an approach to computing the OCR quality of a text document based on the amount of tokens found in a dictionary in relation to all tokens in the text. They suggest to only consider documents with a score higher than 0.7 for NER processing. Furthermore, their computed ratios of a sample of 100 English text documents correlate with the manual ratings of two human OCR raters (p. 101).
- Dinarelli and Rosset (2012) successfully performed three different semi-automated steps for OCR-correction on the corpus of historical French newspapers from the QUAERO[8] project. In a first step, they approached segmentation errors (type II) by concatenating successive lines, if the first line concludes with a dash character. Secondly, they re-tokenized word errors by removing frequent noise from the OCR-ized text, e.g. characters that cannot appear in words (type I). The third step was to manually replace the 300 most frequent out-of-vocabulary (OOV) words with the correct correspondence according to a French dictionary (type III). All steps increased the F1 score of the NER outcome compared to the non-corrected baseline. However, they found the exclusive performance of the second step to be most successful (p. 1267 ff.).
- "*To compensate for the sometimes poor OCR accuracy, an approach was tested where named entities were extracted from the OCR text, and spelling errors deliberately introduced (in the same way they are found in poor OCR results) into the training data. The intention was to cancel out the spelling errors found in the original text. This approach did not produce beneficial results, because the named entity recognizer that was trained on the corrupted training data did produce results that were slightly inferior to those that were obtained without this extra step.*" (Neudecker 2016, p. 4350).

- annotation guidelines (Grouin et al. 2011)
  - Yimam et al. (2014) found that automated pre-tagging of named entities in German texts saved about a fifth of manual annotation time. However, their case study was performed on non OCR-ized texts with an F1 Score around 0.8 for the pre-tagging algorithm (p. 95).

- evaluation measures & outcomes

[8] http://www.quaero.org/

3. Berlin Historical Newspaper Corpus - Clemens, Kai, Vivien

The corpus used in this study is based on the digitized newspaper collection of the Berlin State Library (SBB) and is mostly derived from the *Zeitungsinformationssystem* repository (ZEFYS)[9]. The newspapers were published in Germany in the late 19th and early 20th century. The corpus consists of mostly German texts with standardized orthography (Labusch et al. 2019, p. 3) with a large variation in fonts and layouts. The complete Berlin newspaper text corpus consists of 2,078,127 historical newspaper pages with three newspapers contributing the bulk of the pages (see table 1).

Title	Time span	# of pages	Share in %
Berliner Börsenzeitung	1872-1931	642,480	30.92
Berliner Tageblatt	1877-1939	489,983	23.58
Berliner Volkszeitung	1890-1930	142,403	6.85
Deutsches Nachrichtenbüro	1936-1940	7,429	0.36
Neueste Mittheilungen	1882-1894	1,322	0.06
Norddeutsche Allgemeine Zeitung	1878-1918	120,362	5.79
Provinzial-Correspondenz	1863-1884	1,087	0.05
Teltower Kreisblatt	1856-1896	25,819	1.24
Vossische Zeitung	1857-1917	647,242	31.15

Table 1: Newspapers in the data set for SoNAR (IDH) full text annotations.

The majority of the newspapers employ a variety of blackletter fonts and complex layouts throughout the document, i.e. headers, article paragraphs, illustrations and advertisements all differ in font size and styling and there are typically multiple columns per page. The number of columns can vary and even change mid-page. Certain terms or phrases as well as loan words are sometimes set in italics. Due to the typesetter choosing letters according to available space amongst other criteria, font size can vary throughout one paragraph and in rarer cases, even in one line.

In the SoNAR[10] research project, a project developing a web-based platform for historical network analysis based on different data sources, this corpus was used in order to show relationships between people and events in archival data (e.g. scholarly letters), bibliographic data (e.g. scholarly publications) and public opinion data (e.g. newspapers).

From this corpus, a representative sample of 65 newspaper pages was derived, which was manually annotated over the course of the project. The sample contained xx sentences and xxx tokens. The system-suggested named entities were intellectually checked and corrected with the NEAT annotation tool.

3.1 Optical Character Recognition

The SBB is currently developing a pipeline for OCR and text enrichment based on state-of-the-art deep learning methods (Neudecker, 2019, Rehm et al., 2020). The pipeline starts with image preprocessing (e.g. binarization, i.e. the conversion of all background pixels to white and all foreground content pixels to black) to improve the performance of subsequent steps. Typically, the digitized newspapers held by SBB were derived by scanning microfilm masters, which frequently show uneven brightness and moderate contrast, which both decrease the quality of the page analysis. Then, a layout analysis model[11] based on Convolutional Neural Networks (CNN) trained from historical newspaper ground truth data determines page areas with content and segments them into text/non-text regions and, subsequently, regions into lines. The extracted lines are then fed to a Tesseract v4 OCR system, which

[9] http://zefys.staatsbibliothek-berlin.de/
[10] https://sonar.fh-potsdam.de/
[11] https://github.com/qurator-spk/eynollah

uses a recognition model that was trained on the GT4HistOCR dataset[12] and that delivers already very good performance for blackletter fonts - always provided the line extraction worked well. In addition to the text recognition, Tesseract v4 is also used to perform segmentation of lines into words, since the SBB NEAT annotation tool and - more importantly - the digital collection online presentation environment require the pixel coordinates on a word level for highlighting purposes.

3.2 Named Entity Recognition, Disambiguation and Linking

The OCR full text is then fed into a pipeline for Named Entity Recognition, Disambiguation and Linking that was developed by the SBB in the Qurator project. The Named Entity Recognition and Entity Linking (NER/EL) pipeline tackles two main problems:

- Named Entity Recognition (NER): Identification of passages of the full text where persons, locations, or organizations are mentioned.
- Entity Linking (EL): Relation of the identified passages of the text to a particular real-world entity by Wikidata-ID if such a relation seems to be likely on the basis of the available context information.

3.2.1 Named Entity Recognition

For a detailed report on the SBB-NER-system see (Labusch et al. 2019) or consider the source code of the SBB-NER system together with additional information that is available for download[13].

Input of the NER system is the OCR text. Outcome of the NER is an annotated text where those passages of the text have been marked that mention a person, a location, or an organisation.

In our pipeline, NER is performed by a BERT-model (Devlin et al. 2018). BERT stands for “Bidirectional Encoder Representations from Transformers”. The Transformer (Vaswani et al., 2017) is a deep neural network model architecture that has been shown to perform well in various NLP tasks (Devlin et al. 2018). It is an attention based non-recurrent NN-architecture that can be efficiently trained on very large data collections.

As a starting point, we used a published BERT-model that has been pre-trained by Google on the text material of the largest 104 Wikipedias. As learning tasks they used a masked-sentence problem as well as a next sentence prediction problem (Devlin et al. 2018). These learning tasks do not require human annotated ground-truth but can be derived directly from any given text material.

By use of the same learning tasks, we performed additional pre-training on historical German OCR texts of the digitised collections of the SBB (2,333,647 pages) in order to accustom the BERT-model to error-prone historical German OCR data. Over the last 10 years, the OCR of the digitised collections of the SBB has been obtained by application of a commercial off-the-shelf OCR-software (Abbyy FineReader) to the scanned images in the digitised collections. That software solution is not specially optimized with respect to historical text material, hence the obtained text contains many OCR errors.

After pre-training, we finally trained the BERT-model with respect to the NER task by use of several NER-ground-truth datasets, i.e., ConLL (Tjong et al., 2003), GermEval (Benikova et al., 2014) and Europeana Newspapers’ German data sets (Neudecker, 2016).

[12] https://zenodo.org/record/1344132
[13] https://github.com/qurator-spk/sbb_ner

Cross-validation results show that for NER in historical German texts, the SBB pre-trained model provides decent performance over a wide range of text material of different epochs and sources (Labusch et al., 2019). In particular, these results show that the pre-training on error-prone historical OCR improves the overall NER-results on historical German text material.

3.2.2 Named Entity Disambiguation and Linking

For a detailed report on the SBB-NEL system see (Labusch et al., 2020) or consider the source code of the SBB-NEL system together with additional information that is available for download[14].

After NER has been performed, those passages in the text have been marked that mention a person, a location, or an organisation. In EL, we now relate those marked passages in the text to some real-world person, location, or organisation, if the context of the marked passages of the text provides sufficient evidence. In that case the relation to some corresponding real-word entity is given as a Wikidata-ID. If there are multiple real-world entities that could be related with sufficient probability, the result is a list of Wikidata-IDs sorted by their matching probability.

Our EL system uses a knowledge base that has been derived from Wikipedia. The results we report here have been obtained with a knowledge base where the identification of persons, locations and organisations in the Wikipedia had been performed on the basis of the Wikipedia category structure. Since this approach has been shown to lead to insufficient coverage, in particular for non-German languages (Labusch et al., 2020), we replaced that approach in the meantime by a Wikidata driven method that identifies relevant entities using SPARQL queries. That new approach results in significantly larger knowledge bases for all languages including German.

The Entity Disambiguation proceeds in three stages:

The first stage is the lookup of possible candidates in the knowledge base. The lookup uses an approximate nearest neighbour search in a space of BERT embeddings in order to identify up to 400 entity candidates that could be related to some text passage in question. The BERT embeddings are computed on the basis of the words in the text passage, hence the 400 candidates are selected only according to a word similarity measure that is implicitly encoded in the embedding space. There is not any consideration of the context of the text passages at this point.

In the second stage of the linking process each of the up to 400 selected candidates is evaluated by means of a text comparison. In that text comparison, text passages of the Wikipedia that have been linked by the human Wikipedia authors to a particular candidate entity are compared with the text passages in the given OCR text that are to be entity linked. This text comparison is performed by another BERT model that has been trained to estimate the probability of two given text passages relating to the same entity.

In the third, final stage, the probabilities of the text comparisons together with additional information from the lookup step are fed into the so-called ranking model. The ranking model is a standard random forest model (Ho, 1995) that computes for each candidate the overall probability of being the correct corresponding real-world entity of a given text passage in question. The final outcome for each marked NER-text passage, i.e., a list of possible related real-world entities, is sorted according to the matching probabilities computed by the random forest model.

[14] https://github.com/qurator-spk/sbb_ned

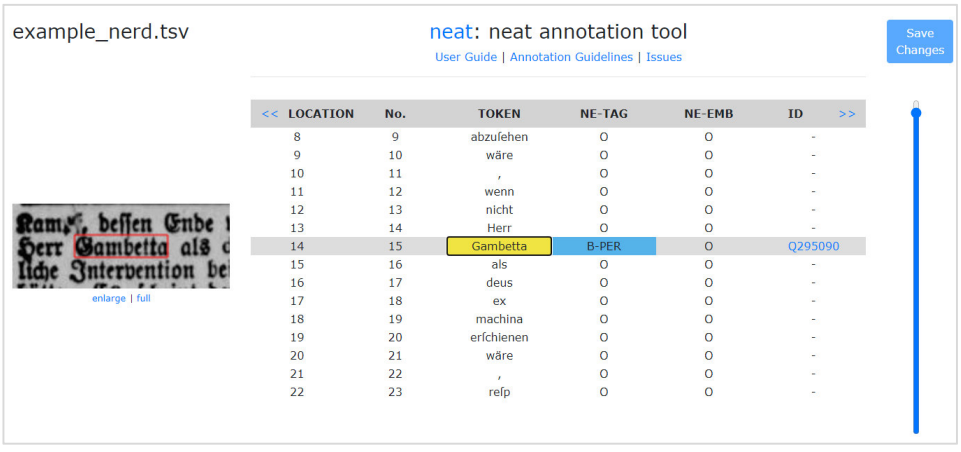
4. The NEAT Annotation Tool - Clemens, Kai

The production of gold standard labeled data for NER/EL training and evaluation requires a suitable annotation software. Various ready available open source tools for this purpose were evaluated by SBB, including BRAT (https://brat.nlplab.org/), WebAnno (https://webanno.github.io/webanno/) and the INL Attestation Tool (https://github.com/INL/AttestationTool). The INCEpTION (https://inception-project.github.io/) tool was not yet widely known at the time the evaluation was done and while the Prodigy (https://prodi.gy/) tool satisfies many of the requirements, it is not very suitable for longer documents, and it requires a paid license which would not have been economical here.

The required features and evaluation criteria where as follows

- works with TSV files based on the GermEval2014 data format (IOB chunking)
- support for large documents (thousands of tokens with no sentence boundaries)
- high annotation speed/ergonomics
- support for embedding image snippets into the annotation environment to support annotators
- support for correction of OCR errors (both token text and segmentation)
- open source (extensibility)
- web browser based
- ease of deployment (no framework/server components, no admin rights required for installation)

While the above mentioned annotation tools typically cover some of these features very well, there was not a single tool that excelled in all relevant aspects. In particular the ability to also correct the text and segmentation had a high priority, since the underlying data for annotation was derived by OCR and thus contains errors, but this was not offered in most tools. Furthermore, as the documents to be annotated are full newspaper pages with lots of articles and long sentences often containing up to 100 words, a tool is required that can deal with such long token sequences. Accordingly, it was decided to create the NEAT annotation tool as a very simple solution specifically for this purpose.



<<	LOCATION	No.	TOKEN	NE-TAG	NE-EMB	ID	>>
	8	9	abzuſehen	O	O	-	
	9	10	wäre	O	O	-	
	10	11	,	O	O	-	
	11	12	wenn	O	O	-	
	12	13	nicht	O	O	-	
	13	14	Herr	O	O	-	
	14	15	Gambetta	B-PER	O	Q295090	
	15	16	als	O	O	-	
	16	17	deus	O	O	-	
	17	18	ex	O	O	-	
	18	19	machina	O	O	-	
	19	20	erſchienen	O	O	-	
	20	21	wäre	O	O	-	
	21	22	,	O	O	-	
	22	23	reſp	O	O	-	

Figure X: Screenshot of the NEAT user interface

NEAT has been implemented as a simple, browser-based HTML and Javascript application. To use it, only a modern web browser with Javascript enabled is required. NEAT operates on a tab-separated-values (TSV) data format that is loosely based on the GermEval2014 data format with IOB chunking.

The user can load a TSV file in the right format and NEAT will render it as a web page with multiple columns that offer different features for editing the content.

The leftmost column “LOCATION” indicates the running number of tokens for the current document, whereas the “No.” signifies the position of a token per sentence (sentence boundaries are marked with a “0”). When clicking the “No.”, a menu opens that allows either merging, splitting or deleting the token. This is helpful to correct errors that occurred during the word segmentation. The next column “TOKEN” holds the token text and can be edited as a text string. A virtual keyboard with a special font allows also entering non-Unicode codepoints such as e.g. from MUFI[15] or OCR-D[16], that frequently occur in historical documents. The next two columns contain the entity tags (which are subsequently color coded) for the surface entity (“NE-TAG”) and also allow the labeling of embedded entities (“NE-EMB”) - to give an example: in the sentence “The President of the United States”, the “President of the United States” is an entity of the type person, while it also includes the location “United States” as an embedded entity. Finally, the last column “ID” can be used for the Wikidata-ID for the surface entity (where available) - and provides this as a clickable link. In order to aid the annotation, NEAT embeds image snippets into the annotation environment through IIIF Image API links in combination with the pixel coordinates for the bounding boxes provided from the OCR output.

All the functionality offered by NEAT can be controlled both with a mouse or with a keyboard, with the latter typically being the more ergonomic option. The code is available as open source from GitHub[17]. A separate Python program[18] is provided for the transformation of OCR results in PAGE-XML format into the TSV data format used by NEAT. The same program can also be used to perform a pre-tagging with NER/EL using the respective tools developed by the SBB and described under 3.2.

5. OCR Correction and NER Annotation - Vivien

This section describes the processes and guidelines for the manual OCR corrections and NER annotations that were performed on the corpus in order to assess the quality of the automatic NER processes with and without OCR corrections. The annotations guidelines[19] were published on GitHub and can be used by other groups. The manual corrections and annotations establish a Gold standard both for the OCR text as well as the identified named entities in the newspaper documents, which will also be released for re-use to the community. For the creation of the best possible ground truth of annotated named entities, we paired the guideline development along with an intensive testing phase on the Berlin newspaper text corpus. This included parallel annotations by two different annotators to assess whether the guidelines ensured homogeneous annotations and continuous discussions about ambiguous cases. Additionally, we introduced the “TODO”-tag in NEAT, which may be used for ambiguous or uncertain tokens in order to support discussion and clarification on the guidelines. This test phase was crucial to ensure the alignment of the guidelines to the historical character of the text corpus.

5.1 OCR Correction

The annotation team corrected characters, words, punctuation marks and other errors in the recognized sentence structure of the text according to a set of agreed upon rules that were included in the

[15] https://folk.uib.no/hnooh/mufi/
[16] https://ocr-d.de/en/gt-guidelines/trans/ocr_d_koordinationsgremium_codierung.html
[17] https://github.com/qurator-spk/neat
[18] https://github.com/qurator-spk/page2tsv
[19] https://github.com/qurator-spk/neat/blob/master/Annotation_Guidelines.pdf

annotation guidelines for the corpus. They were discussed and adapted whenever necessary. The OCR correction was not performed on the entire text, but just on the manually identified named entities. For character or word errors, the spelling of words was adjusted and missing characters or words were added after consulting the snippet or the complete scan of the document and recognizing an entity among the erroneous tokens. Two or more tokens, if incorrectly recognized as one (i.e., a segmentation error), were separated and start-of-sentence-markers were placed or removed according to the sentences recognizable in the scan or snippet. Sentence boundaries (i.e. sectional errors) were only corrected if it interfered with the correct annotation of named entities.
In multiple cases, exceptions had to be documented as correction was not possible or would have decreased the quality of the OCR text. For example, if the correct spelling of a token was indeterminable even after consulting the scan, the spelling of this token would remain untouched. It is important to note that named entity annotation was still possible, however, as long as the entity type of the token was discernible. Neither orthography nor grammar were corrected if it meant opposing the original scan, including suspected grammatical or spelling errors as well as printing or typesetting mistakes. In some cases, the differentiation of characters proved challenging. Some blackletter fonts do not distinguish between the capital I and J, which made it impossible to determine the appropriate correction based on the original, so in these cases the OCR was accepted as valid. The historical blackletter character hyphen ("⸗") was kept according to the original, but hyphenation separating tokens over two or more lines (representing a line break in the original) were corrected to show the word as a single token. Hyphens separating compound words were kept whenever orthographically relevant and otherwise deleted as the tokens were combined into one. While separating punctuation marks from word tokens, the annotation team decided on exceptions for abbreviations and numerals, e.g. in dates. Furthermore, not all special characters and accents could be considered during correction. Next to historical characters like "ſ" and "⸗", the special characters that were corrected in the entities were “Ü”,”ü”,”Ö”,”ö”,”Ä”,”ä” and ”ß” as well as the accents aigu (é), grave (è) and circumflex (ê). Other special characters were left as is (for example, "Bj0rn" was not corrected to "Bjørn" or "Numa" zu "Nuña").
Overall, OCR correction took about 150 minutes per page and added a lot of complexity to the workflow. Great differences in the OCR quality of different newspaper pages meant that some pages needed significantly longer correction efforts. Errors embedded in the original newspaper (e.g. spelling, typesetting, physical damage or discoloration) also slowed down the process.
The corrected OCR text was sent through the NER pipeline again to compare automatic NER with and without OCR correction.

Kommentiert [1]: @zinckjos@hu-berlin.de Dies ist mir noch nicht ganz deutlich. Wurden nicht nur Sachen korrigiert, die direkt an einer Named Entity dran waren? Ich habe den Satz davor dazugeschrieben. Und anstatt "missing" tokens habe ich erroneous tokens geschrieben. Ist dies so korrekt? _zinckjos@hu-berlin.de zugewiesen_

Kommentiert [2]: Wir haben den OCR-Text gelesen und alle offensichtlichen Entitäten korrigiert. Gab es lediglich den Verdacht auf eine Entität, haben wir das Schriftbild im Scan geprüft. Wenn in der OCR "Mün n" gestanden hätte, hätten wir im Scan gecheckt ob es "München" ist oder evtl. eine andere Entität und nur dann eine Korrektur/bzw. Vervollständigung vorgenommen. Bei Rückfragen einfach auf diesen Kommentar antworten. :)

Kommentiert [3]: Nochmal per Logbuch kalkulieren zur Sicherheit.

5.2 NER Annotation

The system-recognized named entities were manually checked and corrected according to the annotation guidelines that were developed in the project. The named entity types considered for annotation of the Berlin newspaper corpus are the following six classes:

- Persons (PER), named entities referring to definite individuals or collectives (e.g. families);
- Organizations (ORG), named entities that refer to political bodies, companies and the like;
- Locations (LOC), named entities referring to a politically or geographically defined place;
- Conferences (CONF), named entities referring to uniquely named gatherings of individuals on a certain pre-defined scholarly or other topic, goal or shared purpose as well as a pre-defined ending point;
- Events (EVT), named entities referring to uniquely identifiable events apart from conferences;
- Works and expressions (WORK), named entities referring to titled human creations (e.g. paintings, films, literary works).

They were chosen based on the available entity classes in the German Integrated Authority File[20], which is one of the knowledge bases considered for entity linking in the SoNAR project. While persons, locations, and organizations are long established classes in NER (Ling et al., p. 323), by introducing the additional classes CONF, EVT, WORK to the NER annotation process, we hoped to assess the potential of entity types beyond the traditional classes. For this, an important preliminary consideration was a qualitative review of a small sample of each newspaper, which allowed for estimations on the expected content.

-- sentences on first and emb and merged levels ---

Any entity token may contain multiple entities or a single entity belonging to multiple entity classes, e.g. entities like hospitals [ORG] which were named after a person [PER] or Stonehenge [LOC and WORK]. The annotators included a section in the guidelines on deciding which entity or entity class will be annotated on the first level (TAG) and which one is an embedded entity (EMB), an entity on the second level: "If one entity marks the entire (group of) token(s) while the other entity marks only parts of it/them or derives from it/them, the latter is the second level entity. If more than one named entity is embedded in another named entity, the annotator chooses which entity is to be marked on the second level by evaluating the nesting levels: Subject/object of the sentence is the first level entity, while its direct attribute is the second level entity. The third level component is to be left out."

The project-specific annotation guidelines were developed iteratively and are available for download[21]. An initial version was set up based on former work with historical texts (Ehrmann et al. 2020b; Fort et al. 2009; Grouin et al. 2011; Reiter 2017; Reznicek 2013; Rosset et al. 2011) and similar projects[22]. The guidelines contain 15 general annotation rules and between 2-8 specific rules for each individual entity type. For the CONF and WORK entities, the guidelines of the German Integrated Authority File and the international bibliographic cataloging standard RDA are considered. Specific German language examples are used for disambiguation. The rules do not only disambiguate overlapping cases, but also determine how much of the surrounding tokens are to be included in the named entity annotation.

Historical named entity annotation remains a challenge, even when done manually. For example, annotators may lack historical context, such as contemporary phrases that were common quotes from then-popular works of literature or persons. Sometimes, context was missed because only one newspaper page was considered at any given time, for example when abbreviations or nicknames were resolved on another page. In order to correctly identify organizations, expert historical knowledge is required on the political, postal or financial systems as well as the internal structures of large organizations (like railway companies, stock exchanges or government departments). Ethical conflicts arose when problematic phrasing or propagandic determiners were annotated (e.g. Nazi propaganda slogans), since even the acknowledgement in an annotation might reinforce possibly systemically racist or sexist ideas. Mixed entity types pose annotation challenges for any type of text, for example, when the same name is used for a location and an organization (e.g. Kreml) or when organizations are named after people or families are named after locations (especially if that family acts as business in itself, like aristocratic houses). This is also true for entities fitting into multiple types equally well, like scientific or trade expeditions. The appropriate detail level for entity recognition was a particular problem for locations and organizations. In the newspaper corpus, for example, locations are often referred to by the exact street address of a building (e.g. a hospital branch in a Berlin neighborhood), not the broader organizational entity that historical researchers might search for.

5.3 Inter-Annotator Agreement

Agreement measures help to show problems in entity annotation and were used to support the annotation guideline development. Inter-annotator agreement describes the level of consensus in the annotation decisions of more than one annotator on the same document sample (Nowak and Rüger 2010, p. 558). Two annotators simultaneously annotated 14 of the OCRized newspaper pages so that metrics could be calculated for the validation of the annotation guidelines.

We distinguish between the agreement over all tokens in a page (ALL), which includes the null decision on whether a token constitutes an entity, and the agreement over those selected tokens, which were

Kommentiert [4]: @zinckjos@hu-berlin.de Liebe Frau Zinck, bitte hier kurz beschreiben, was dies ist, sonst ist die Erklärung im inter-annotation Kapitel nicht verständlich.
zinckjos@hu-berlin.de zugewiesen

Kommentiert [5]: Ich bin noch nicht ganz sicher, was genau gemeint ist: soll ich eine Erklärung zu first level entities und embedded entities geben? Was bedeutet dann merged dann in dem Fall - mixed types?

Kommentiert [6]: @zinckjos@hu-berlin.de
Ja genau! Wenn Sie nicht wissen, was merged ist, dann weiß ich es auch nicht. ;) Taucht in der Tabelle von Sina auf, ich frage sie. Sie schrieb im Text: both levels combined (MERGED). Das besprechen wir vielleicht erst in 5.3, also erklären Sie bitte first und emb, ja?

Kommentiert [7]: Table: https://docs.google.com/spreadsheets/d/1oFfqhFzBx3A8-K7CxpHX5As9labfAHKYVWnFnJ6E170/edit?usp=sharing

[20] https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html
[21] https://github.com/qurator-spk/neat/blob/master/Annotation_Guidelines.pdf
[22] Examples: QUAERO (http://www.quaero.org/) and the Impresso project (https://impresso-project.ch/).

identified as an entity by at least one annotator (SEL). As figure 4 illustrates, the share of identified entity tokens differs significantly over the newspaper pages, which also impacts agreement rates, due to the heterogeneous content within pages as well as the corresponding newspaper sections. A page-filling stock exchange index, listing numerous organizations contains a very high share of mentioned entities, whereas a page containing advertisements might not. In the case of page 3, for example, which contains over 50% entity tokens, the newspaper section contained a list of people and no continuous text.

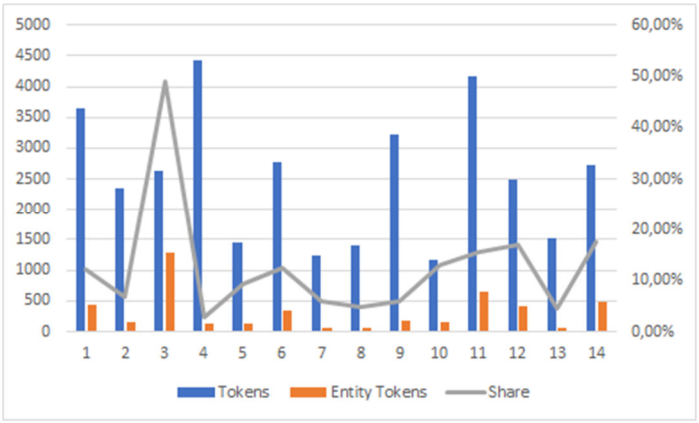


Figure 4: Relations of entity tokens and all tokens of each document annotated by 2 annotators (left side shows absolute number of tokens, right side reflects the percentage of entity tokens out of all tokens).

We further measured the agreement on first (TAG) and second (EMB) level, and both levels combined (MERGED).

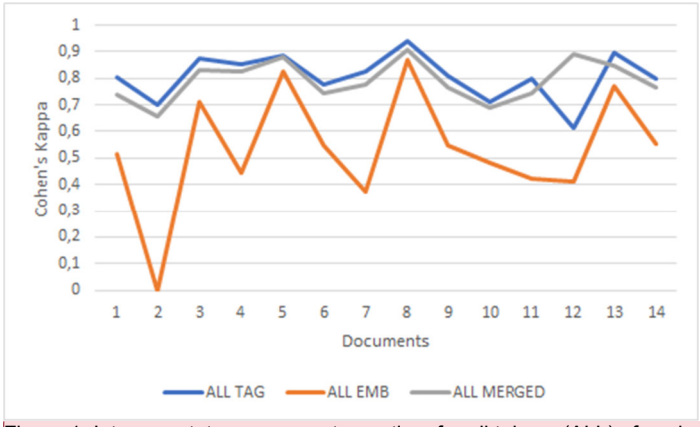


Figure 1: Inter annotator agreement over time for all tokens (ALL) of each document on the first level (TAG), second level (EMB) and both combined (MERGED).

Both annotators started off with a rather high rate of agreement, which can be seen both as an indicator for a solid bracketing effect of the guidelines as well as thorough briefing and training sessions before starting annotations. There was especially low agreement on embedded entities. As the guidelines developed along with the annotation process, there is a slight upward trend of agreements over all tokens, which is due to the high amount of 0-decisions over all tokens (meaning none of the annotators marked a token as (part of) an entity). If a mutual 0-decision is included in the agreement measures, a low rate of found entities in a document works in favor of the agreement rate.

Kommentiert [8]: Die Abb. sind hier interessant, sollten wir für die finale Publikation aber rausnehmen und nur verbal erklären. Spannend sidn die Schwankungen pro Dokument.

Kommentiert [9]: check, whether that phrase is common

More importantly, there is a slight decrease of agreement rates over the entity tokens. This is due to impactful repetitive errors within single documents, e.g. in the outliers document 2 and 12, which had an exceptionally low agreement rate. For document 12, for example, this was caused by different assumptions on the annotation of embedded locations if an entity refers to a historical place, which also has a current correspondent. These errors had a large impact on single documents, but could immediately be eliminated by additions to the guidelines.

Kommentiert [10]: das ist wichtig, dass wir diese Analyse zum Korrigieren der Guidelines benutzt haben

6. Sample Newspaper Corpus for Evaluation - Hannes

The current ground truth corpus holds 19 OCRized pages from all newspapers in the corpus, consisting of 51,992 tokens. In the 19 pages of the ground truth corpus, a total of 3602 entities were annotated, which is an average of 189 entities per newspaper page (redundant mentions of the same entity included). 11,8% of the 51,992 token were annotated as (part of) a named entity (6135 token) at an average length of 1,7 token per entity.

Kommentiert [11]: Leider ist es wg. Flexion und Koreferenz sehr schwierig, die types (also die Anzahl der von den markierten Token referenzierten verschiedenen Entitäten) automatisiert festzustellen, bis auf eine manuelle Auswertung der 3602 Entitäten fällt mir keine Möglichkeit dazu ein, außer wir warten noch die NEL-Annotation ab, dann kann man es anhand der Q-Nummern aus Wikidata zumindest für jene Entitäten abgleichen, die eine Verlinkung gekriegt haben. Die anderen müssten weiterhin manuell disambiguiert und ausgezählt werden.

Entity type	Total	First level (TAG)	Second level (EMB)
Location (LOC)	1408	990	418
Person (PER)	1002	828	174
Organization (ORG)	943	743	200
Work (WORK)	130	117	13
Event (EVT)	116	109	7
Conference (CONF)	3	3	0
Total	3602	2790	812

Table 2: Entities in the ground truth corpus.

As one might expect, the traditional entity types LOC (39,09%), PER (27,82%), and ORG (26,18%) were most frequently found, whereas the other classes WORK (3,61%), EVT (3,22%), and especially CONF (0,08%) may be considered a supporting addition, which might be of importance to researchers in historical network analysis.

Complete table:
https://docs.google.com/spreadsheets/d/1EDzFuRbYxHup4y1SSK887TuCJ8go578gPTyli4Dfp-E/edit?usp=sharing

7. Results - Hannes

7.1 Overall NER Quality

RQ1 How good is BERT NER on historical newspaper material in German?
comparison with other studies

Kommentiert [12]: Based on which configuration? We don't have intellectual evaluation data for the last roundtrip. Only the older version. but that we can do.

As in Dinaralli/Rosset 2012 (p. 1267 f.):

- Spurious tokens
- unreliable punctuation

7.2 NER Quality by Entity Type

RQ2 Does NER quality differ by entity type?

- explanations and hypotheses why (from annotation experience)

7.3 NER Quality and OCR Corrections

RQ3 Does NER quality differ if we OCR correct or not?

- how many OCR corrections?
- Corrections on the Berlin newspaper corpus concentrate on error types I, II, and IV. They exclusively concern errors occurring in named entities.

X.Y Selection of pages for annotation

- Is it feasible/beneficial to ex-/include pages with advertisements?
- Impact of segmentation errors on OCR/annotation
- ...

8. Conclusion - Vivien

- what did we find
- continued annotation for bigger ground truth
- NER quality may be good, but what about NEL quality if the entity is misspelled? → next paper

9. Acknowledgements

Acknowledgements. This research is part of the research projects SoNAR (IDH) funded by the DFG – German Research Foundation (grant no. 414792379) and QURATOR funded by the German Federal Ministry of Education and Research (BMBF) (Unternehmen Region, Wachstumskern, grant no. 03WKDA1A).

10. References

Alex, Beatrice; Burns, John (2014): Estimating and rating the quality of optically character recognised text. In: Apostolos Antonacopoulos und Klaus U. Schulz (Hg.): Digital Access to Textual Cultural Heritage. DATeCH 2014 : conference proceedings : Madrid, May 19-20, 2014. the First International Conference. Madrid, Spain, 5/19/2014 - 5/20/2014. New York, NY, USA: ACM (ICPS), p. 97–102.

Darina Benikova; Chris Biemann; Max Kisselew; Sebastian Pad´o (2014): GermEval 2014 Named Entity Recognition: Companion paper. In; Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition, Hildesheim, Germany, pages 104–112.

Chiron, Guillaume; Doucet, Antoine; Coustaty, Mickaël; Visani, Muriel; Moreux, Jean-Philippe (2017): Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information. In: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), p. 1–4.

Dinarelli, Marco; Rosset, Sophie (2012): Tree-Structured Named Entity Recognition on OCR Data: Analysis, Processing and Results. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey: European Language Resources

Kommentiert [13]: see above

Kommentiert [14]: this is the trickiest part because I have to align all kinds of different versions. my test programs so far aren't very good at that.

Kommentiert [15]: I'll try to have some numbers about the impact of advertisements on the quality. those won't be statistically significant but the whole study isn't really. see my point below about the thousands of pages needed for that

Kommentiert [16]: the selection is tricky. because for equal coverage of the corpus we would have needed to annotate a few thousand pages.

Association (ELRA), p. 1266–1272. http://www.lrec-conf.org/proceedings/lrec2012/pdf/1046_Paper.pdf.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ehrmann, Maud; Romanello, Matteo; Bircher, Stefan; Clematide, Simon (2020a): Introducing the CLEF 2020 HIPE Shared Task: Named Entity Recognition and Linking on Historical Newspapers. In: Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva und Flávio Martins (Hg.): ADVANCES IN INFORMATION RETRIEVAL. 42nd european conference on ir research, Bd. 12036. [S.l.]: Springer (Lecture notes in computer science), p. 524–532. https://link.springer.com/chapter/10.1007/978-3-030-45442-5_68

Ehrmann, Maud; Watter, Camille; Romanello, Matteo; Clematide, Simon; Flückiger, Alex (2020b, January 10). Impresso Named Entity Annotation Guidelines (Version 2.2.0). Zenodo. http://doi.org/10.5281/zenodo.3604227

Ehrmann, Maud; Romanello, Matteo; Clematide, Simon; Ströbel, Phillip Benjamin; Barman, Raphaël (2020c): Language Resources for Historical Newspapers: the Impresso Collection. In: Proceedings of The 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, S. 958–968. Online verfügbar unter https://www.aclweb.org/anthology/2020.lrec-1.121.

Fort, Karën; Ehrmann, Maud; Nazarenko, Adeline (2009): Towards a Methodology for Named Entities Annotation. In: Proceedings of the Third Linguistic Annotation Workshop (LAW III), p. 142-145. https://www.aclweb.org/anthology/W09-3025.pdf

Galibert, Olivier; Rosset, Sophie; Grouin, Cyril; Zweigenbaum, Pierre; Quintard, Ludovic (2012): Extended Named Entities Annotation on OCRed Documents: From Corpus Constitution to Evaluation Campaign. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey: European Language Resources Association (ELRA), p. 3126–3131. Online verfügbar unter http://www.lrec-conf.org/proceedings/lrec2012/pdf/343_Paper.pdf.

Geyken, Alexander, Haaf, Susanne, Jurish, Bryan, Schulz, Matthias, Thomas, Christian, & Wiegand, Frank (2012). TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv. *Jahrbuch für Computerphilologie*, *9*.

Grouin, Cyril, Rosset, Sophie, Zweigenbaum, Pierre, Fort, Karën, Galibert, Olivier, & Quintard, Ludovic (2011). Proposal for an Extension of Traditional Named Entities: From Guidelines to Evaluation, an Overview. *ACL HLT 2011*, 92. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.207.5409&rep=rep1&type=pdf#page=104

Grover, Claire; Givon, Sharon; Tobin, Richard; Ball, Julian (2008): Named Entity Recognition for Digitised Historical Texts. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). Marrakech, Morocco: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/pdf/342_paper.pdf

Hamdi, Ahmed; Jean-Caurant, Axel; Sidere, Nicolas; Coustaty, Mickael; Doucet, Antoine (2019): An Analysis of the Performance of Named Entity Recognition over OCRed Documents. In: Maria Bonn (Hg.): 2019 ACM/IEEE Joint Conference on Digital Libraries. JCDL 2019 : proceedings : 2-6 June 2019, Urbana-Champaign, Illinois. 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). Champaign, IL, USA, 6/2/2019 - 6/6/2019. Piscataway, NJ: IEEE, p. 333–334.

https://zenodo.org/record/3243344/files/JCDL_2019_An%20Analysis%20of%20the%20Performance%20of%20Named%20Entity%20Recognition.pdf

Ho, Tin Kam (1995): Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. Vol. 1. IEEE.

Kaplan, Frédéric; Di Lenardo, Isabella (2017): Big Data of the Past. In: Front. Digit. Humanit. 4. DOI: 10.3389/fdigh.2017.00012.

Kettunen, Kimmo; Ruokolainen, Teemu (2017): Names, Right or Wrong: Named Entities in an OCRed Historical Finnish Newspaper Collection. In: Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2017. Göttingen, Germany: ACM Press, p. 181–186. http://dl.acm.org/citation.cfm?doid=3078081.3078084

Kugler, Anna (2018): Automatisierte Volltexterschließung von Retrodigitalisaten am Beispiel historischer Zeitungen. 33-54 Seiten / Perspektive Bibliothek, Bd. 7, Nr. 1 (2018). DOI: 10.11588/PB.2018.1.48394.

Labusch, Kai; Neudecker, Clemens; Zellhöfer, David (2019): BERT for Named Entity Recognition in Contemporary and Historic German. In: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019).
https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/KONVENS2019_paper_4.pdf

Labusch, Kai; Neudecker, Clemens (2020): Named Entity Disambiguation and Linking Historic Newspaper OCR with BERT. In: CLEF 2020.
http://ceur-ws.org/Vol-2696/paper_163.pdf

Landis, J. Richard; Koch, Gary G. (1977): The Measurement of Observer Agreement for Categorical Data. In: Biometrics 33 (1), p. 159–174. DOI: 10.2307/2529310.

Ling, Xiao; Singh, Sameer; Weld, Daniel p. (2015): Design Challenges for Entity Linking. In: Transactions of the Association for Computational Linguistics 3 (1), p. 315–328. DOI: 10.1162/tacl_a_00141.

Linhares Pontes, Elvys; Hamdi, Ahmed; Sidere, Nicolas; Doucet, Antoine (2019): Impact of OCR Quality on Named Entity Linking. In: Adam Jatowt, Akira Maeda und Sue Yeon Syn (Hg.): Digital libraries at the crossroads of digital information for the future. 21st International Conference on Asia-Pacific Digital Libraries, ICADL 2019, Kuala Lumpur, Malaysia, November 4-7, 2019, Proceedings, Bd. 11853. Cham: Springer (LNCS sublibrary. SL 3, Information systems and applications, incl. Internet/Web, and HCI, v. 11853), p. 102–115.

Lopresti, Daniel (2009): Optical character recognition errors and their effects on natural language processing. In: IJDAR 12 (3), p. 141–151. DOI: 10.1007/s10032-009-0094-8.
http://www.cse.lehigh.edu/~lopresti/tmp/AND08journal.pdf

Neudecker, Clemens (2016): An Open Corpus for Named Entity Recognition in Historic Newspapers. In: Proceedings of the 10th Language Resources and Evaluation Conference. Portorož, Slovenia. Online verfügbar unter https://www.aclweb.org/anthology/L16-1689.

Neudecker, Clemens; Antonacopoulos, Apostolos (2016): Making Europe's Historical Newspapers Searchable. In: DAS 2016. 12th IAPR International Workshop on Document Analysis Systems : 11-14 April 2016, Santorini, Greece : proceedings. 2016 12th IAPR Workshop on Document Analysis Systems (DAS). Santorini, Greece, 4/11/2016 - 4/14/2016. IAPR International Workshop on Document Analysis Systems.

Neudecker, Clemens; Baierer, Konstantin; Federbusch, Maria; Würzner, Kay-Michael; Boenig, Matthias; Herrmann, Elisa; Hartmann, Volker: OCR-D: An end-to-end open-source OCR framework for historical documents, in: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage, Brüssel 09.05.2019, S. 53–58. 2019.

Neudecker, Clemens; Wilms, Lotte; Faber, Willem Jan; van Veen, Theo: Large scale refinement of digital historical newspapers with named entity recognition. In: Proceedings of the IFLA 2014 Newspaper Section Satellite Meeting, 13-14 August 2014, Geneva, Switzerland, pp. 1-15, 2014.

Nowak, Stefanie; Rüger, Stefan (2010): How reliable are annotations via crowdsourcing. In: James Z. Wang, Nozha Boujemaa, Nuria Oliver Ramirez und Apostol Natsev (Hg.): Proceedings of the international conference on Multimedia information retrieval - MIR '10. the international conference. Philadelphia, Pennsylvania, USA, 29.03.2010 - 31.03.2010. New York, New York, USA: ACM Press, p. 557. http://www.computerphilologie.de/jg09/geykenetal.pdf

Packer, Thomas L.; Lutes, Joshua F.; Stewart, Aaron P.; Embley, David W.; Ringger, Eric K.; Seppi, Kevin D.; Jensen, Lee S. (2010): Extracting person names from diverse and noisy OCR text. In: Roberto Basili (Hg.): Proceedings of the fourth workshop on Analytics for noisy unstructured text data. the fourth workshop. Toronto, ON, Canada, 10/26/2010 - 10/26/2010. ACM Special Interest Group on Information Retrieval; ACM Special Interest Group on Hypertext, Hypermedia, and Web; ACM Special Interest Group on Knowledge Discovery in Data. New York, NY: ACM, p. 19. https://www.deg.byu.edu/papers/Ancestry_NAACL_HLT_Paper.pdf

Rehm, G., Bourgonje, P., Hegele, S., Kintzel, F., Schneider, J. M., Ostendorff, M., ... & Heine, F. (2020). QURATOR: Innovative Technologies for Content and Data Curation. In: Adrian Paschke, Clemens Neudecker, Georg Rehm, Jamal Al Qundus, Lydia Pintscher (editor). Proceedings of QURATOR 2020 -- The conference for intelligent content solutions. Conference on Digital Curation Technologies (QURATOR-2020) January 20-21 Berlin, Germany CEUR Workshop Proceedings 2/2020. http://ceur-ws.org/Vol-2535/paper_17.pdf

Reiter, Nils (2017): How to Develop Annotation Guidelines. Blog post. Available at: https://sharedtasksinthedh.github.io/2017/10/01/howto-annotation/

Reznicek, Marc (2013): Linguistische Annotation von Nichtstandardvarietäten —Guidelines und „Best Practices". Guidelines NER. Version 1.5. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/nosta-d/nosta-d-ner-1.5

Rodriquez, Kepa Joseba, Bryant, Mike, Blanke, Tobias, & Luszczynska, Magdalena (2012). Comparison of named entity recognition tools for raw OCR text. In *Konvens* (pp. 410-414). http://www.oegai.at/konvens2012/proceedings/60_rodriquez12w/60_rodriquez12w.pdf

Rosset, Sophie; Grouin, Cyril; Zweigenbaum, Pierre (2011): Entités Nommées Structurées : guide d'annotation Quaero (Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum), Technical report. Available at: http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf

Erik F.; Tjong Kim Sang; Fien De Meulder (2003): Introduction to the conll-2003 shared task:Language-independent named entity recognition. In: Proceedings of the Seventh Conference on NaturalLanguage Learning at HLT-NAACL 2003 - Volume 4,CONLL '03, pages 142–147, Stroudsburg, PA, USA.Association for Computational Linguistics.

Traub, Myriam C.; van Ossenbruggen, Jacco; Hardman, Lynda (2015): Impact Analysis of OCR Quality on Research Tasks in Digital Archives. In: Sarantos Kapidakis, Cezary Mazurek und Marcin Werla (Hg.): Research and advanced technology for digital libraries. 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015 : proceedings, Bd. 9316. Cham: Springer International Publishing (Lecture notes in computer science, 9316), p. 252–263. https://link.springer.com/chapter/10.1007/978-3-319-24592-8_19

van Strien, Daniel, Beelen, Kaspar, Ardanuy, Mariona Coll., Hosseini, Kasra, McGillivray, Barbara, & Colavizza, Giovanni (2020). Assessing the Impact of OCR Quality on Downstream NLP Tasks. In *ICAART (1)* (pp. 484-496). https://www.staff.universiteitleiden.nl/binaries/content/assets/governance-and-global-affairs/isga/artidigh_2020_7_cr.pdf

Vaswani, Ashish; Noam Shazeer; Niki Parmar; Jakob Uszkoreit; Llion Jones; Aidan N. Gomez; Lukasz Kaiser; Illia Polosukhin (2017): Attention is all you need. In: arXiv preprint arXiv:1706.03762 (2017).

Yimam, Seid Muhie; Eckard de Castilho, Richard; Gurevych, Iryna; Biemann, Chris (2014): Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore, MD, USA, p. 91–96.

Galibert, Olivier, Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum, und Ludovic Quintard. 2012. „Extended Named Entity Annotation on OCRed Documents: From Corpus Constitution to Evaluation Campaign", 6.