How Scalable is Quality Assessment of Text Recognition? A Combination of Ground Truth and Confidence Scores

Michał Bubula[1], Konstantin Baierer[1], Jörg Lehmann[1], Clemens Neudecker[1], Vahid Rezanezhad[1], and Doris Škarić[1]

[1] Department for Information and Data Management, Berlin State Library, Berlin, Germany

Abstract

Vast amounts of historical documents are being digitized with subsequent optical character recognition (OCR), but the quality assessment of the results is challenging for larger quantities. Ground Truth-based evaluation requires sufficient and representative data that are expensive to create. Following recent work, we investigate whether confidence scores automatically provided by text recognition systems can serve as a proxy. Based on an analysis of the relationship between word error rates and word confidence scores for several OCR engines, we find that the latter can serve as a useful indicator of OCR quality. In a second step, we explore the scalability and reliability of combining Ground Truth and confidence scores for quality assessment of text recognition in several experiments on a heterogeneous dataset comprising almost 5 million pages of historical documents from 1456–2000. The deeper analysis of the evaluation results provides insights into typical issues for OCR of historical documents, suggesting potential directions for future work.

Keywords: optical character recognition, evaluation, confidence scores, word error rates, ground truth, historical document analysis

1 Introduction

The primary method for evaluating the quality of optical character recognition (OCR) involves comparing outputs with manually transcribed reference texts, commonly referred to as Ground Truth (GT) [16]. This approach facilitates the quantification of error rates by computing the edit distance (Levenshtein distance) between the recognized text and the correct transcription. However, the production of GT data entails substantial effort, typically limiting such evaluations to relatively small samples. Consequently, the validity and generalizability of the resulting quality assessments diminish as the sample size decreases relative to the overall volume and heterogeneity of the OCR-processed material. On the other hand, mass digitization in libraries and archives has led to millions of pages of full-texts derived by OCR models. OCR quality has significant impact on downstream tasks [12; 23; 26] such as full-text search or natural language processing such as named entity recognition [5; 8], text and data mining or text analysis [10]. The source materials spanning hundreds of years contain significant variations with regard to the visual properties, typefaces, text lengths and publication dates that characterize them (see Figure 1).

Accordingly, representativeness by sampling is hard to achieve against the background of a large collection comprising of highly heterogeneous source materials. This raises the question: what alternative methods for OCR quality assessment exist that are scalable but also reliable?

Michał Bubula, Konstantin Baierer, Jörg Lehmann, Clemens Neudecker, Vahid Rezanezhad, and Doris Škarić. "How Scalable is Quality Assessment of Text Recognition? A Combination of Ground Truth and Confidence Scores." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 1267–1291. https://doi.org/10.63744/GR59c1iXu6Wj.

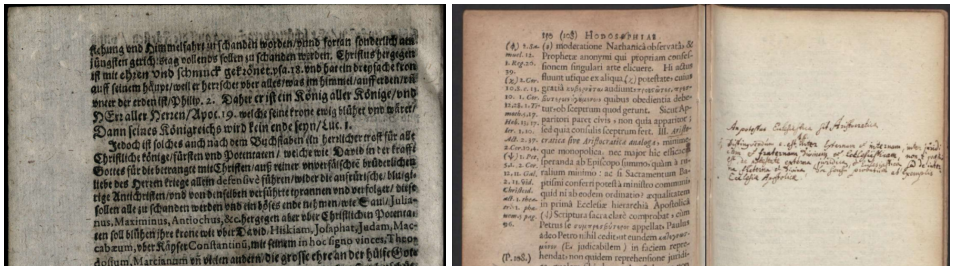

Figure 1: Examples of two historical documents, dated 1643 (left) and 1666 (right), illustrating the challenges posed by diversity in visual quality, printed and handwritten letters, and layouts.

2 Related Work

The research community has long sought approaches to OCR evaluation that do not rely on GT data. One such method leverages lexicons to identify out-of-vocabulary words, which are then flagged as potential errors [1; 7]. While theoretically promising, this approach is often insufficient for historical documents, which frequently exhibit extensive orthographic variation over time, and for which period matching and machine readable historical dictionaries are scarce or even unavailable. Another method proposed in [19] uses character-level N-gram models, which must also be trained per-language, exhibiting similar problems as dictionary-based methods in the context of historical spelling variants. A SVM-based approach is pursued in [28], which requires a manually classified garbage/non-garbage token training set. There are also combinations of the above; e.g. [21] explores combining a dictionary approach with N-grams or SVMs.

More recently, confidence scores—internal metrics generated by OCR engines that serve as a form of automated self-assessment for each recognized word—have been investigated as an alternative strategy [4], including the use of LLMs like BERT [9]. Pseudo-perplexity scores provided by BERT are explored by [24] who show that it can be a meaningful metric for quality estimation especially when appropriate lexical resources are not available. But how realistic and reliable are these confidence scores across larger collections with high variation of document types? If such scores are correlated with quality metrics derived from small samples of GT, it becomes possible to measure the discrepancy between the OCR engine's internal confidence computation and the actual accuracy determined by comparison with GT. If this discrepancy remains relatively stable across different inputs, the actual quality of OCR output may be approximated by appropriately calibrating the confidence scores.

The primary objective of our experiment is to gather empirical insights through statistical analyses and case-based examinations, and to identify the document characteristics that positively or negatively influence recognition performance within the applied OCR workflow. These insights aim to inform a strategy whereby documents that exhibit similar features are grouped and processed using OCR models and workflows tailored towards the specific features of each group, thereby optimizing recognition quality. The overarching goal is to gather knowledge about configurations where document features and OCR engine or model choices match so that they can be generalized to other similar documents, given that individual case-by-case OCR processing is infeasible due to the quantity of the source material.

3 Data

The analysis of the relationship between word confidence scores (WC) and word error rates (WER, see [17]) is conducted using the following GT datasets: an unreleased dataset provided by the Vlaamse Erfgoedbibliotheken (VEB), which includes 75 pages of historical Belgian newspapers,

a dataset from the OCR-D project (OCR-D-GT).[1] and a dataset from the Berlin State Library (OCR-D-GT-VD-SBB) consisting of 348 pages.[2] The models employed for the subsequent evaluations include: a CNN-RNN model for Eynollah;[3] the `deep3_lsh4` model for Calamari;[4] the `Reichsanzeiger` model for Kraken;[5] and the `german_print`[6] and `deu_frak`[7] models for Tesseract.[8]

The approximately 47,000 historical documents processed with OCR were previously determined to be single-column German-language works from VD projects[9] based on their metadata and a custom image analysis algorithm. The majority of the works is printed in Fraktur. They can all be accessed via the Digitized Collections of Berlin State Library.[10]

4 WER–-WC Relationship Analysis

4.1 Evaluation of Methods

Prior to the evaluation of the OCR results for the 47,000 documents, it was necessary to assess whether the confidence scores provided by the OCR engine used (in this case, Tesseract) can indeed serve as reliable indicators of recognition accuracy. Our approach to verification involves examining the relationship between word confidence scores and word error rates (see Appendix B for a detailed explanation of the latter).

We hypothesize that these metrics exhibit a consistent relationship: a low WER should correspond to high confidence scores, reflecting accurate recognition, whereas a high WER, indicating numerous errors, should be associated with lower confidence values. To examine this correlation, we employed several smaller datasets with available GT and also processed them using various OCR engines, including Eynollah [20], Calamari [27], Kraken [13], and Tesseract [22].

The dataset is partitioned into two subsets: a training set (blue crosses in Figure 2), comprising 70% of the data, used to develop a predictive model, and a test set (orange crosses), comprising the remaining 30% and utilized to evaluate the model's generalizability to unseen data. To quantitatively characterize this relationship, two regression models are applied: a linear regression model and a second-degree (quadratic) polynomial regression model [4].

The linear regression model is defined as

$$\text{wer(wc)} = \beta_1 \cdot \text{wc} + \beta_0,$$

where β_0 represents the intercept and β_1 denotes the slope of the regression line. In addition to the linear model, we also consider a second-degree polynomial regression to account for potential nonlinear effects. The quadratic regression model is given by

$$\text{wer(wc)} = \beta_2 \cdot \text{wc}^2 + \beta_1 \cdot \text{wc} + \beta_0,$$

[1] `https://github.com/OCR-D/gt_structure_text/releases/tag/v1.5.0`
[2] `https://doi.org/10.5281/zenodo.17395956`
[3] `https://zenodo.org/records/17194824`
[4] `https://github.com/Calamari-OCR/calamari_models_experimental/tree/main/deep3_lsh4`
[5] `https://ub-backup.bib.uni-mannheim.de/~stweil/tesstrain/kraken/reichsanzeiger-gt/reichsanzeiger_best.mlmodel`
[6] `https://ub-backup.bib.uni-mannheim.de/~stweil/tesstrain/german_print/`
[7] `https://github.com/tesseract-ocr/tessdata/blob/main/deu_frak.traineddata`
[8] `german_print` and `deu_frak` models are not directly comparable in confidence output but internally consistent, cf. K.
[9] VD is the abbreviation for *"Verzeichnis der im deutschsprachigen Raum erschienenen Drucke"*, i.e. all books of the 16th, 17th and 18th centuries printed in German-language countries. `https://vd16.de`, `http://www.vd17.de/`, `vd18.de`.
[10] `https://digital.staatsbibliothek-berlin.de/`

where β_2 is the coefficient of the squared term and captures any curvature in the relationship between word count and error rate. Here, β_1 remains the linear coefficient and β_0 is the intercept as before.

These models provide interpretable estimates of the relationship between OCR confidence scores and actual recognition performance. The quadratic model, in particular, allows for modeling more complex, nonlinear patterns that may occur at the extremes of the confidence scale.

To assess predictive performance, we report the coefficient of determination (R^2), which indicates the proportion of variance explained by the model, and the mean squared error (MSE), reflecting average prediction error. Pearson's correlation coefficient (r) measures the strength of linear association between predicted and observed values, while Spearman's rank correlation coefficient (ρ) captures monotonic relationships based on the relative ordering of the data, offering robustness to non-linear trends and outliers. For each regression coefficient β_j, we report the corresponding p-value to assess statistical significance. Low p-values (typically < 0.05) suggest that the predictor contributes meaningfully to the model. Regression plots include shaded 95% confidence intervals to visualize the uncertainty around predictions; narrower intervals indicate higher reliability. A detailed discussion of these metrics and their interpretation is provided in Appendix C.

4.2 Evaluation of Tesseract

Tesseract [22] represents one of the most widely established [15; 18] OCR engines, making it a relevant choice for this study. The experiment shown in Figure 2 is performed using the `german_print` model on the OCR-D-GT and VEB datasets, and the `deu_frak` model on the OCR-D-GT-VD-SBB dataset.

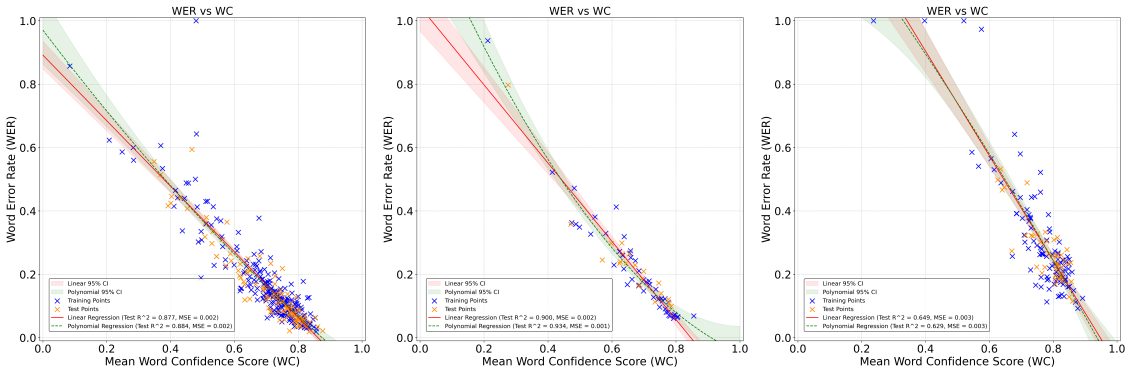

Figure 2: Relationship between WER and WC evaluated with Tesseract on OCR-D-GT (left), VEB (middle), and OCR-D-GT-VD-SBB (right).

Table 1 summarizes model performance metrics including correlation coefficients, R^2, and MSE for both linear and polynomial regressions across the three datasets. Table 2 provides the corresponding regression coefficients and p-values, indicating the strength and significance of each model term.

Dataset	r	ρ	Lin. R^2	Lin. MSE	Poly. R^2	Poly. MSE
OCR-D-GT	-0.942	-0.914	0.877	0.002	0.884	0.002
VEB	-0.949	-0.909	0.900	0.002	0.934	0.001
OCR-D-GT-VD-SBB	-0.838	-0.674	0.649	0.003	0.629	0.003

Table 1: Performance metrics for the OCR-D-GT, VEB, and OCR-D-GT-VD-SBB datasets.

Dataset	Lin. β_1 (p)	Lin. β_0 (p)	Poly. β_2 (p)	Poly. β_1 (p)	Poly. β_0 (p)
OCR-D-GT	-1.034 (.000)	0.892 (.000)	0.248 (.115)	-1.326 (.000)	0.971 (.000)
VEB	-1.234 (.000)	1.045 (.000)	0.948 (.001)	-2.346 (.000)	1.349 (.000)
OCR-D-GT-VD-SBB	-1.652 (.000)	1.565 (.000)	-0.267 (.538)	-1.309 (.022)	1.461 (.000)

Table 2: Regression coefficients with their corresponding p-values for the OCR-D-GT, VEB, and OCR-D-GT-VD-SBB datasets.

The regression analyses reveal strong and consistent relationships between WC scores and WER across all three datasets. The linear regression models generally perform well, with high R^2 coefficients, low MSE, and strong correlation coefficients (r and ρ), indicating that confidence scores are meaningful predictors of recognition performance.

Among the datasets, the VEB dataset shows the highest predictive performance, with a linear R^2 of 0.900 and a slightly improved polynomial R^2 of 0.934, along with the lowest MSE (0.001). Correlation coefficients are also very strong ($r = -0.949$, $\rho = -0.909$), and the polynomial model yields a significant quadratic term (β_2, $p = .001$), suggesting that a mild nonlinear effect may improve predictions in this case.

The OCR-D-GT dataset also performs strongly, with both linear and polynomial models achieving R^2 values above 0.87. The linear model already captures most of the variance ($r = -0.942$), and the marginal gain from the polynomial fit is not supported by a statistically significant quadratic term ($p = .115$). This supports the use of the simpler linear model for practical interpretation.

While the OCR-D-GT-VD-SBB dataset shows somewhat lower R^2 values (0.649 for the linear model), the relationship between confidence scores and WER clearly remains present. Both linear and rank correlations ($r = -0.838$, $\rho = -0.674$) indicate a meaningful trend, and the linear slope is highly significant ($p = .000$). The reduced model fit may be partially explained by the use of a different OCR model for this dataset, which may introduce differences in confidence calibration or recognition behavior. Variations in the layout, typography, or quality of the source material may also contribute to the observed variation in model performance. Still, the polynomial model offers no substantial improvement and includes a non-significant quadratic term ($p = .538$), further supporting the appropriateness of a linear interpretation.

Across all datasets, the linear slope coefficient (β_1) is consistently negative and statistically significant, confirming the expected inverse relationship: as OCR confidence increases, WER tends to decrease. The quadratic terms, by contrast, are generally not significant and provide limited practical benefit.

To sum up, linear regression offers a robust, interpretable, and statistically well-supported model for predicting OCR error rates from confidence scores. Although polynomial models may offer slight improvements in select cases, these gains are minor and come at the cost of increased model complexity. The consistency of linear trends across varied datasets, including under different Tesseract models, supports the linear model as the most practical and reliable approach for analyzing OCR performance. This evaluation using Eynollah, Calamari, and Kraken is presented in Appendix D.

5 Confidence Score Evaluation on Historical Documents

5.1 Evaluation of the Complete Dataset

The analysis utilizes input data comprising confidence scores assigned by the OCR engine (Tesseract with the `deu_frak` model) to each recognized word and textline within a page of a work, identified by PPN (Pica Production Number), an identifier assigned to records within the German

library network Gemeinsamer Bibliotheksverbund (GBV). These scores range from 0 to 1, with 1 indicating perfect recognition accuracy. Figure 3 presents WC in the first row and TC in the second. Within each row, the left-most plot shows the distribution of mean confidence values, while the subsequent plots present the corresponding standard deviations, computed directly from the original scores per PPN. A low standard deviation indicates consistent recognition certainty across a PPN, whereas higher values point to greater variability, potentially due to image quality, alternating typefaces, or layout challenges (see Appendix A).

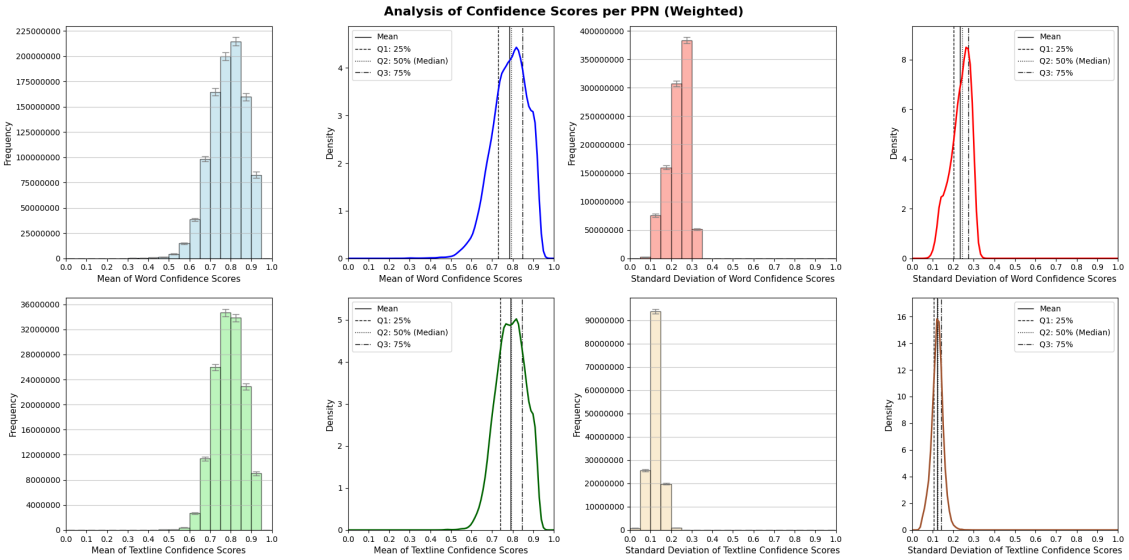

Figure 3: The evaluation of the complete dataset of historical documents from 1456–2000.

To mitigate potential bias caused by PPNs with limited text content, a weighted analysis was applied (see Appendix E). For word-level analysis, each PPN was weighted by its total word count, and for textline-level analysis, by the number of textlines. To assess the precision of our estimates, we compute standard errors using an effective sample size that accounts for the uneven distribution of weights across documents.

Two types of visualizations are employed: bar charts and density plots. The bar charts present weighted histograms with confidence score intervals of 0.5, chosen to balance detail and smoothness. The y-axis represents the total number of words or textlines per confidence interval, with error bars indicating uncertainty from weighted sampling. Both word- and textline-level histograms reveal a concentration of confidence scores within the 0.75–0.9 range, indicating high certainty of the OCR engine. For a more refined view of the data's structure, kernel density estimation is applied, yielding smooth probability distributions for WC and TC scores with means of 0.785 ± 0.004 and 0.791 ± 0.004, respectively.

A detailed analysis of OCR confidence in relation to document size is provided in Appendices F and G. Confidence scores increase with document length up to a moderate size, stabilizing around 0.80–0.82 for documents of 52,500–165,000 words, 6,750–15,750 textlines, or 200–360 pages. Medium-length documents exhibit the most consistent OCR performance, with narrow error margins indicating robust estimates. In contrast, very short or very long documents show greater variability and lower confidence, often due to minimal text, complex layouts, degraded scans, or heterogeneous content. These findings underscore document length as a key factor influencing OCR quality and guide the identification of reliable subsets for downstream processing.

5.2 Evaluation by Centuries

The dataset covers a broad time range from the 15th to the 20th century. Table 3 summarizes the number of documents per century alongside the mean WC and TC values, each accompanied by their standard error of the mean (SEM).

Century	Number of publications	Mean WC $\pm$ SEM	Mean TC $\pm$ SEM
15th	225	0.665 ± 0.045	0.687 ± 0.046
16th	6565	0.700 ± 0.009	0.719 ± 0.009
17th	11993	0.706 ± 0.006	0.726 ± 0.007
18th	21533	0.793 ± 0.005	0.800 ± 0.005
19th	4382	0.854 ± 0.013	0.855 ± 0.013
20th	1786	0.883 ± 0.021	0.882 ± 0.021

Table 3: Evaluation of publication dates across centuries.

The statistical analyses for each century demonstrate a clear temporal trend, indicating a progressive improvement in recognition accuracy over time. In the earliest period examined, the 15th century, the mean WC and TC values are relatively low (0.665 ± 0.045 and 0.687 ± 0.046, respectively), reflecting the challenges associated with early printed materials. This is consistent with the irregularities and lack of standardization characteristic of early typographic practices. As printing techniques evolved during the 16th and 17th centuries, both confidence metrics show gradual improvement, with the mean WC rising to 0.706 and TC to 0.726 by the 17th century.

A notable increase in OCR accuracy is observed beginning with the 18th century, where both WC and TC exceed 0.79. This improvement correlates with increased typographic standardization, the invention of lithography and higher print quality during the Enlightenment period. The 19th century continues this upward trajectory, reaching mean WC and TC values of approximately 0.854 and 0.855, respectively, likely facilitated by the widespread adoption of mechanical typesetting techniques and the steam-powered rotary printing press. The highest confidence values are attained in the 20th century, with both WC and TC averaging around 0.883, indicative of modern printing practices being disproportionally reflected in the training data for the model used.

An exemplary analysis of the 17th-century data is discussed in Appendix J. A comprehensive overview of the genres represented in the dataset is provided in Appendix H. To further understand the variability in OCR performance across different types of material, we also conduct a detailed evaluation of subgenres within the dataset, presented in Appendix I. These analyses enable us to assess how genre-specific characteristics may influence recognition accuracy, offering insights into the strengths and limitations of the evaluated OCR models.

6 Conclusion and Further Work

This analysis focused on two aspects: first, examining the correlation between WER and WC, and second, using the latter to evaluate a large dataset of historical documents. The WER–-WC relationship was analyzed across different GT datasets using OCR engines such as Tesseract, Eynollah, Calamari, and Kraken. Linear and polynomial regressions revealed a clear negative correlation: low WER values correspond to high WC scores, and vice versa. Subsequently, Tesseract's `deu_frak` model was applied to approximately five million pages of historical documents from the 15th to the 20th century. Tesseract was chosen for its prevalence in OCR workflows and the practical interpretability of its linear WER–-WC correlation.

An evaluation of nearly 47,000 historical documents examined which characteristics influence recognition performance in standard OCR workflows. Publication date emerged as a key factor:

typefaces and typographic conventions from the 18th century onward are more effectively recognized by the Fraktur-based model than those from earlier periods. Our analysis further reveals a clear upward trend in OCR quality over the centuries, reflecting advances in printing technology and standardization. These insights can serve as a basis for developing grouping methods that categorize documents by features such as document length, genre, and publication date, enabling the application of customized OCR workflows to enhance overall full-text quality.

The question of how genres, publication dates, and related features are represented across different models is undoubtedly highly relevant. However, this topic extends beyond the scope of the present study and should be addressed in future research. A comprehensive investigation of this nature would benefit from broad systematic comparisons and carefully established GT for each genre and century, providing a valuable foundation for more detailed and fine-grained analyses.

It should be noted that although the present analysis was limited to a regression of two (or three) variables, future work could employ hierarchical multiple regression to incorporate additional covariates. Potential parameters such as document length, genre, publication date, and variations in typeface or language have already been identified in this paper. While [3] has taken initial steps in this direction, future analyses should explicitly define independent and dependent variables and account for potential moderators, such as binarization, that may alter the strength or direction of their relationships.

The present analysis identifies several areas for improving OCR workflows in large-scale digitization of historical documents. Key aspects include image preprocessing, layout analysis [6], and the selection of suitable models and engines for text recognition. When binarization reduces image quality and impairs recognition or layout detection, applying precise cropping focused on text regions can markedly enhance results. For documents with complex typography, such as marginal notes, ornaments, tables, or intricate layouts, more advanced layout analysis methods are needed. Conventional, fast layout analysis tools, such as those provided by Tesseract, may be insufficient to segment such structures accurately, whereas more sophisticated tools can substantially improve segmentation and layout understanding.

Quality assessment of text recognition can be scaled by combination of GT data and OCR confidence scores. The correlation between WER and WC supports reliable quality estimation across diverse datasets and OCR engines. This approach enables large-scale evaluation of historical collections, revealing trends related to document length, temporal variation, and typefaces. Scalability depends not only on computational efficiency, but also on selecting OCR models suited to specific historical contexts. As libraries continue to digitize vast collections, this method offers a practical framework for improving full-text quality at scale. Its effectiveness, however, remains limited by the availability and representativeness of GT datasets, which vary across periods and genres. Nevertheless, efforts to expand and diversify GT resources hold the potential for more detailed comparative analyses across historical contexts.

The source code used for this analysis is published on GitHub.[11]

References

[1] Alex, Beatrice and Burns, John. “Estimating and rating the quality of optically character recognised text”. In: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*. 2014, pp. 97–102. DOI: 10.1145/2595188.2595214.

[2] Correia, Sergio and Luck, Stephan. “Digitizing historical balance sheet data: A practitioner’s guide”. In: *Explorations in Economic History 87* (Jan. 2023), p. 101475. DOI: 10.1016/j.eeh.2022.101475.

[11] https://github.com/qurator-spk/sbb_ocr_conf_eval

[3] Cuper, Mirjam. “Examining a Multi Layered Approach for Classification of OCR Quality without Ground Truth”. In: *DH Benelux Journal. The Humanities in a Digital World 4* (July 2022), pp. 42–59. DOI: 10.17613/03DS-9973.

[4] Cuper, Mirjam, Dongen, Corine van, and Koster, Tineke. “Unraveling confidence: examining confidence scores as proxy for OCR quality”. In: *International Conference on Document Analysis and Recognition*. Springer. 2023, pp. 104–120. DOI: 10.1007/978-3-031-41734-4_7.

[5] Ehrmann, Maud, Romanello, Matteo, Flückiger, Alex, and Clematide, Simon. “Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers”. Publisher: CEUR-WS. Sept. 2020. DOI: 10.5167/UZH-200192.

[6] Fleischhacker, David, Kern, Roman, and Göderle, Wolfgang. “Enhancing OCR in historical documents with complex layouts through machine learning”. In: *International Journal on Digital Libraries 26*, no. 1 (Mar. 2025). Publisher: Springer Science and Business Media LLC. ISSN: 1432-5012, 1432-1300. DOI: 10.1007/s00799-025-00413-z.

[7] Gupta, Anshul, Gutierrez-Osuna, Ricardo, Christy, Matthew, Capitanu, Boris, Auvil, Loretta, Grumbach, Liz, Furuta, Richard, and Mandell, Laura. “Automatic assessment of OCR quality in historical documents”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 1. 2015. DOI: 10.1609/aaai.v29i1.9487.

[8] Hamdi, Ahmed, Jean-Caurant, Axel, Sidere, Nicolas, Coustaty, Mickael, and Doucet, Antoine. “An Analysis of the Performance of Named Entity Recognition over OCRed Documents”. In: *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Champaign, IL, USA: IEEE, June 2019, pp. 333–334. DOI: 10.1109/JCDL.2019.00057.

[9] Hemmer, Arthur, Coustaty, Mickaël, Bartolo, Nicola, and Ogier, Jean-Marc. “Confidence-Aware Document OCR Error Detection”. In: *Document Analysis Systems*, ed. by Giorgos Sfikas and George Retsinas. Vol. 14994. Series Title: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2024, pp. 213–228. DOI: 10.1007/978-3-031-70442-0_13.

[10] Hill, Mark J and Hengchen, Simon. “Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study”. In: *Digital Scholarship in the Humanities 34*, no. 4 (2019), pp. 825–843. DOI: 10.1093/llc/fqz024.

[11] Holley, Rose. “How Good Can It Get?: Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs”. In: *D-Lib Magazine 15*, no. 3/4 (Mar. 2009). Publisher: CNRI Acct. ISSN: 1082-9873. DOI: 10.1045/march2009-holley.

[12] Jerele, Ines, Erjavec, Tomaž, Pokorn, Daša, and Kavčič-Čolić, Alenka. “Optical Character Recognition of Historical Texts: End-User Focused Research for Slovenian Books and Newspapers from the 18th and 19th Century”. In: *Review of the National Center for Digitization* , no. 21 (2012). Publisher: Faculty of Mathematics, pp. 117–126. URL: http://eudml.org/doc/254555.

[13] Kiessling, Benjamin. “Version 5 of the Kraken ATR Engine for the Humanities”. In: *Document Analysis and Recognition – ICDAR 2025: 19th International Conference, Wuhan, China, September 16–21, 2025, Proceedings, Part III*. Wuhan, China: Springer - Verlag, 2025, pp. 443–458. DOI: 10.1007/978-3-032-04624-6_26. URL: https://doi.org/10.1007/978-3-032-04624-6_26.

[14] Kirchner, Felix, Dittrich, Marco, Beckenbauer, Phillip, and Nöth, Maximilian. “OCR bei Inkunabeln – Offizinspezifischer Ansatz der Universitätsbibliothek Würzburg”. In: *ABI Technik 36*, no. 3 (Sept. 2016), pp. 178–188. ISSN: 2191-4664, 0720-6763. DOI: 10.1515/abitech-2016-0036.

[15] Koistinen, Mika, Kettunen, Kimmo, and Pääkkönen, Tuula. “Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing”. In: *Proceedings of the 21st Nordic Conference on Computational Linguistics*, ed. by Jörg Tiedemann and Nina Tahmasebi. Gothenburg, Sweden: Association for Computational Linguistics, May 2017, pp. 277–283. URL: https://aclanthology.org/W17-0238/.

[16] Neudecker, Clemens, Baierer, Konstantin, Gerber, Mike, Clausner, Christian, Antonacopoulos, Apostolos, and Pletschacher, Stefan. “A survey of OCR evaluation tools and metrics”. In: *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*. 2021, pp. 13–18. DOI: 10.1145/3476887.3476888.

[17] Neudecker, Clemens, Zaczynska, Karolina, Baierer, Konstantin, Rehm, Georg, Gerber, Mike, and Schneider, Julián Moreno. “Methoden und Metriken zur Messung von OCR-Qualität für die Kuratierung von Daten und Metadaten”. In: *Qualität in der Inhaltserschließung*, ed. by Michael Franke-Maier, Anna Kasprzik, Andreas Ledl, and Hans Schürmann. De Gruyter, Sept. 2021, pp. 137–166. DOI: 10.1515/9783110691597-009.

[18] Novotný, Vít. “When Tesseract Does It Alone: Optical Character Recognition of Medieval Texts”. In: *RASLAN – Recent Advances in Slavonic Natural Processing* (Dec. 2020). Publisher: Tribun EU, pp. 3–22. ISSN: 2336-4289. URL: https://nlp.fi.muni.cz/raslan/raslan20.pdf.

[19] Popat, Ashok C. “A panlingual anomalous text detector”. In: *Proceedings of the 9th ACM symposium on Document engineering*. 2009, pp. 201–204. DOI: 10.1145/1600193.1600237.

[20] Rezanezhad, Vahid, Baierer, Konstantin, Gerber, Mike, Labusch, Kai, and Neudecker, Clemens. “Document Layout Analysis with Deep Learning and Heuristics”. In: *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*. HIP ’23. San Jose, CA, USA: Association for Computing Machinery, 2023, pp. 73–78. DOI: 10.1145/3604951.3605513.

[21] Sankaran, Naveen and Jawahar, CV. “Error detection in highly inflectional languages”. In: *2013 12th International Conference on Document Analysis and Recognition*. IEEE. 2013, pp. 1135–1139. DOI: 10.1109/ICDAR.2013.230.

[22] Smith, Ray. “An overview of the Tesseract OCR engine”. In: *Ninth international conference on document analysis and recognition (ICDAR 2007)*. Vol. 2. IEEE. 2007, pp. 629–633. DOI: 10.1109/ICDAR.2007.4376991.

[23] Strien, Daniel van, Beelen, Kaspar, Ardanuy, Mariona, Hosseini, Kasra, McGillivray, Barbara, and Colavizza, Giovanni. “Assessing the Impact of OCR Quality on Downstream NLP Tasks”. In: *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*. Valletta, Malta: SCITEPRESS - Science and Technology Publications, 2020, pp. 484–496. DOI: 10.5220/0009169004840496.

[24] Ströbel, Phillip Benjamin, Volk, Martin, Clematide, Simon, Schwitter, Raphael, Hodel, Tobias, and Schoch, David. “Evaluation of HTR models without Ground Truth Material”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, June 2022, pp. 4395–4404. URL: https://aclanthology.org/2022.lrec-1.467/.

[25] Tanner, Simon, Muñoz, Trevor, and Ros, Pich Hemy. “Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library’s 19th Century Online Newspaper Archive”. In: *D-Lib Magazine 15*, no. 7/8 (July 2009). Publisher: CNRI Acct. ISSN: 1082-9873. DOI: 10.1045/july2009-munoz.

[26] Traub, Myriam C. “Impact Analysis of OCR Quality on Research Tasks in Digital Archives”. In: *Lecture Notes in Computer Science*, ed. by Jacco Van Ossenbruggen and Lynda Hardman. ISSN: 0302-9743, 1611-3349. Cham: Springer International Publishing, 2015, pp. 252–263. DOI: 10.1007/978-3-319-24592-8_19.

[27] Wick, Christoph, Reul, Christian, and Puppe, Frank. “Calamari – A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition”. 2020. URL: https://dhq.digitalhumanities.org/vol/14/2/000451/000451.html.

[28] Wudtke, Richard, Ringlstetter, Christoph, and Schulz, Klaus U. “Recognizing garbage in OCR output on historical documents”. In: *Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data*. 2011, pp. 1–6. DOI: 10.1145/2034617.2034626.

A Examples of Challenging Cases

While [11], [25], and [12] conducted a quality assessment of OCR accuracy in historical newspapers, [2] in historical micro-data, [14] on incunabula, and [6] worked on printed serial source published during a long time frame, a comprehensive evaluation of OCR quality across a broad corpus spanning multiple centuries, document lengths, genres, complex layouts and typefaces has, to the best of our knowledge, not yet been conducted.

Using 17th century examples as a case study, we aim to illustrate the challenges associated with the diverse range of issues typically encountered in historical documents. In addition to Figure 1 shown in Section 1 depicting two problematic documents, dated 1643[12] (left) and 1666[13] (right), Figure 4 displays three works exemplifying these complexities and highlighting the difficulties arising while analyzing such materials.

From left to right, the first is *Andechtige Gebet/ Gesenge und Collecten/ auff alle Tage in der Wochen* (1605),[14] the second is *Continens Epistolarum Festivalium pericopas* by Christoph Dauderstadt (1656),[15] and the third is *Hodosophia Christiana seu Theologia Positiva* by Johann Conrad Dannhauer (1666).[16]

The main challenges are dark, blurred characters and translucent printing, specks and pages with black margins, heterogeneous documents with complex layouts, marginal notes and ornaments, changing typefaces (Fraktur and Latin) and languages, warped pages with curved textlines, or having very little actual text such as on pages with portraits (with the name of the sitter), city views (with legends) or sheet music (with lines of lyrics).

A detailed case study of the 17th-century documents is presented in Appendix J.

[12] https://digital.staatsbibliothek-berlin.de/werkansicht?PPN=PPN1040867766&PHYSID=PHYS_0094&DMDID=DMDLOG_0003

[13] https://digital.staatsbibliothek-berlin.de/werkansicht?PPN=PPN856676292&view=picture-double&PHYSID=PHYS_0182&DMDID=DMDLOG_0001

[14] https://digital.staatsbibliothek-berlin.de/werkansicht?PPN=PPN717767116&PHYSID=PHYS_0297&DMDID=DMDLOG_0001

[15] https://digital.staatsbibliothek-berlin.de/werkansicht?PPN=PPN788451650&PHYSID=PHYS_0474&DMDID=DMDLOG_0001

[16] https://digital.staatsbibliothek-berlin.de/werkansicht?PPN=PPN856677698&PHYSID=PHYS_0091&DMDID=DMDLOG_0001

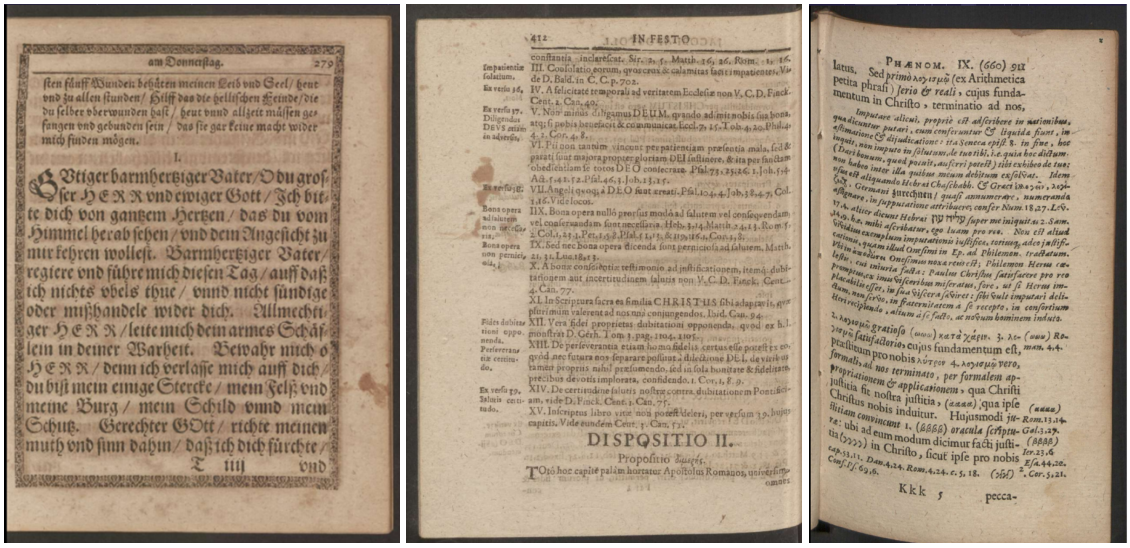

Figure 4: Examples of historical documents from the 17th century.

B Word Error Rates

The word error rate (WER) is a widely used metric for quantifying the accuracy of optical character recognition (OCR) systems, calculated in analogy to the character error rate (CER; see [9], p. 6); the latter is valued as an established measure for quality assessment [17]. WER represents the proportion of recognition errors at the word level relative to a reference transcription referred to as the ground truth. Formally, it is computed by counting the total number of word-level errors—comprising substitutions, deletions, and insertions—and dividing this sum by the total number of words in the ground truth transcription.

Mathematically, the WER is expressed as:

$$\text{WER} = \frac{S + D + I}{N} \tag{1}$$

where:

- S is the number of substitutions (e.g., recognizing “Maus” instead of “Haus”),
- D is the number of deletions (missed words),
- I is the number of insertions (extra words inserted into the recognition output),
- N is the total number of words in the ground truth transcription.

This metric provides an interpretable measure of OCR performance: lower WER values indicate higher accuracy, with zero representing a perfect recognition.

A key aspect of evaluating OCR performance is establishing what constitutes acceptable or good recognition quality. According to established standards in the field, a character error rate (CER) below 10% is generally considered indicative of high-quality OCR (see [3], p. 46). Similarly, a word error rate (WER) of 20% or below is often regarded as a benchmark for acceptable OCR accuracy, as noted by [23]. These values are derived from the tasks which the OCR results serve. Though 10% CER cannot be considered as truly satisfying, it can be seen as providing acceptable results for full-text search. A word error rate of 20% means that every fifth word is not correctly recognized. As a result on their research on improving OCR metrics in serial publications, [6] present final CER scores of 4,33% and final WER scores of 21,1%. However, it has to be

noted that the transparent provision of CER and WER scores in historical document OCR is still a desideratum [26].

Achieving these thresholds suggests that the OCR system produces results with minimal errors, making the transcriptions reliable for further analysis or digital use. Conversely, higher error rates may require additional post-processing or review to ensure data quality. These benchmarks serve as useful guidelines for assessing the performance of OCR systems in various applications, particularly when working with historical materials where recognition challenges are more pronounced.

C Regression Model Evaluation Metrics

To evaluate the predictive performance of the models, several statistical metrics are employed, each capturing different aspects of model quality.

First, the coefficient of determination (R^2) is used to quantify how well the model explains the variability in the observed data. It is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where y_i and $\hat{y}_i$ denote the observed and predicted values, respectively, $\bar{y}$ is the mean of the observed values, and n is the number of observations. An R^2 value of 1 indicates a perfect fit, while a value of 0 suggests that the model explains none of the variance in the response variable.

The mean squared error (MSE) complements R^2 by quantifying the average squared difference between the predicted and observed values:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Lower MSE values indicate better predictive accuracy, with greater penalties assigned to larger errors due to the squaring of residuals.

In addition to these error-based metrics, we compute the Pearson correlation coefficient (r) to assess the strength and direction of the linear relationship between predicted and observed values:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}},$$

where $\bar{y}$ and $\bar{\hat{y}}$ are the means of the observed and predicted values, respectively. Pearson's r ranges from -1 to 1, with values close to 1 indicating a strong positive linear association.

To capture potential monotonic but non-linear relationships, we also include the Spearman rank correlation coefficient (ρ), which evaluates the strength of a monotonic association based on ranked data. It is defined as

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

where d_i is the difference between the ranks of y_i and $\hat{y}_i$. By relying on ranks instead of raw values, Spearman's ρ is more robust to outliers and is well-suited for detecting non-linear yet consistently ordered patterns.

For each regression coefficient β_j ($j = 0, 1, 2$), we report the associated p-value to test the null hypothesis $H_0 : \beta_j = 0$, i.e., that the coefficient has no effect. A low p-value (typically < 0.05) suggests that the corresponding predictor significantly contributes to the model, providing evidence against the null hypothesis.

The regression plots presented in the analysis include shaded regions around the fitted lines representing 95% confidence intervals. These intervals illustrate the uncertainty associated with

the predicted regression line: if the experiment was repeated under similar conditions, we expect the true regression line to lie within the shaded bounds 95% of the time. Narrower intervals indicate greater certainty and higher reliability of the model in those regions.

D WER–WC Evaluation with Eynollah, Calamari and Kraken

D.1 Eynollah

The experiments shown in this section are performed using Eynollah [20] (see Figure 5).

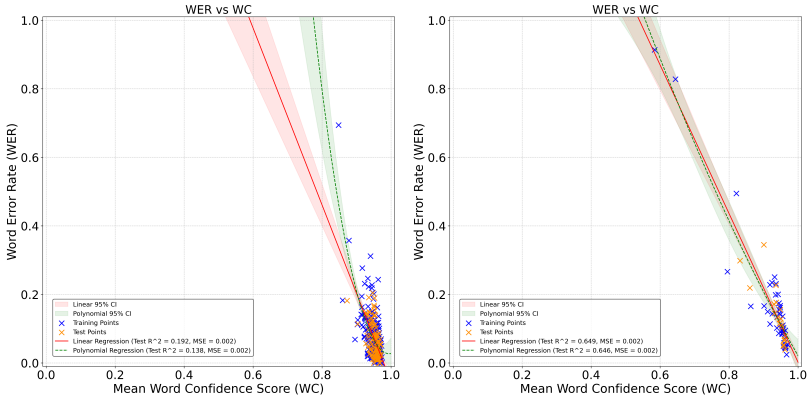

Figure 5: Relationship between WER and WC evaluated with Eynollah on the OCR-D-GT (left) and VEB (right) datasets.

Dataset	r	ρ	Lin. R^2	Lin. MSE	Poly. R^2	Poly. MSE
OCR-D-GT	-0.502	-0.487	0.192	0.002	0.138	0.002
VEB	-0.812	-0.905	0.649	0.002	0.646	0.002

Table 4: Performance metrics for the OCR-D-GT and VEB datasets in the case of Eynollah.

Dataset	Lin. β_1 (p)	Lin. β_0 (p)	Poly. β_2 (p)	Poly. β_1 (p)	Poly. β_0 (p)
OCR-D-GT	-2.581 (.000)	2.526 (.000)	20.307 (.000)	-40.399 (.000)	20.119 (.000)
VEB	-2.169 (.000)	2.171 (.000)	1.010 (.359)	-3.785 (.036)	2.796 (.000)

Table 5: Regression coefficients with their corresponding p-values for the OCR-D-GT and VEB datasets in the case of Eynollah.

The results, shown in Tables 4 and 5, indicate that, for both datasets, the linear regression model generally performs on par or better than the polynomial model. For the VEB dataset, the linear model achieves a high R^2 value of 0.649, nearly identical to the polynomial model (0.646), with both models yielding the same mean squared error (0.002). For the OCR-D-GT dataset, the linear model also slightly outperforms the polynomial model in terms of R^2 (0.192 vs. 0.138), with identical MSE. Notably, the R^2 values for both models are substantially higher on the VEB dataset than on the OCR-D-GT dataset, indicating a much stronger relationship between the variables in the VEB data. Additionally, the p-values for the polynomial term in the VEB dataset are not significant, suggesting that the added complexity of the polynomial model does not provide a meaningful improvement. Overall, the linear regression model is the preferred choice for both datasets.

D.2 Calamari

In the following subsection, we employ Calamari [27] with the `deep3_lsh4` model (see Figure 6).

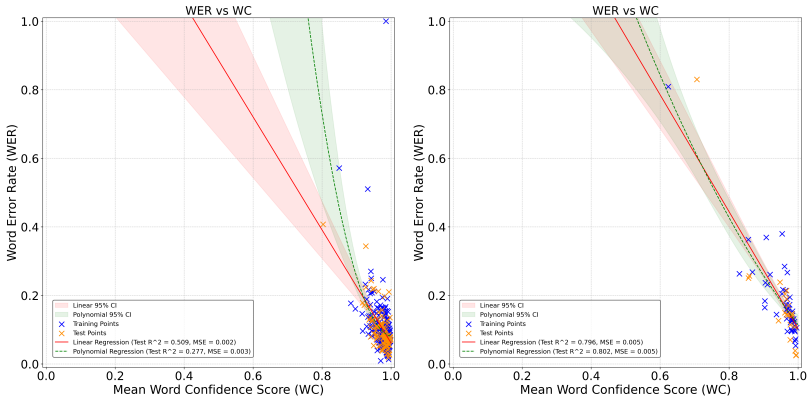

Figure 6: Relationship between WER and WC evaluated with Calamari on the OCR-D-GT (left) and VEB (right) datasets.

Dataset	r	ρ	Lin. R^2	Lin. MSE	Poly. R^2	Poly. MSE
OCR-D-GT	-0.722	-0.551	0.509	0.002	0.277	0.003
VEB	-0.922	-0.814	0.796	0.005	0.802	0.005

Table 6: Performance metrics for the OCR-D-GT and VEB datasets in the case of Calamari.

Dataset	Lin. β_1 (p)	Lin. β_0 (p)	Poly. β_2 (p)	Poly. β_1 (p)	Poly. β_0 (p)
OCR-D-GT	-1.649 (.000)	1.710 (.000)	15.047 (.007)	-30.334 (.004)	15.367 (.002)
VEB	-1.706 (.000)	1.809 (.000)	1.296 (.217)	-3.896 (.032)	2.714 (.001)

Table 7: Regression coefficients with their corresponding p-values for the OCR-D-GT and VEB datasets in the case of Calamari.

Tables 6 and 7 summarize the regression analysis for Calamari across both datasets. In the VEB dataset, both linear and polynomial models yield high R^2 values (0.796 and 0.802, respectively) and identical mean squared errors, indicating that either model captures the relationship between variables effectively. For the OCR-D-GT dataset, the linear regression model demonstrates a clear advantage, with a notably higher R^2 (0.509) and lower MSE compared to the polynomial model. The lack of statistical significance for the polynomial terms in the VEB data further suggests that increasing model complexity does not enhance predictive power. The consistently higher R^2 values observed for the VEB dataset point to a stronger association in that subset. In summary, linear regression is sufficient for both datasets, and compared to Eynollah, Calamari achieves substantially higher R^2 values, particularly on the OCR-D-GT dataset, indicating a better overall model fit.

D.3 Kraken

Finally, we evaluate Kraken [13] with the `Reichsanzeiger` model (see Figure 7).

Tables 8 and 9 present the regression results for Kraken across both datasets. For the OCR-D-GT dataset, the polynomial regression model outperforms the linear model, achieving a higher

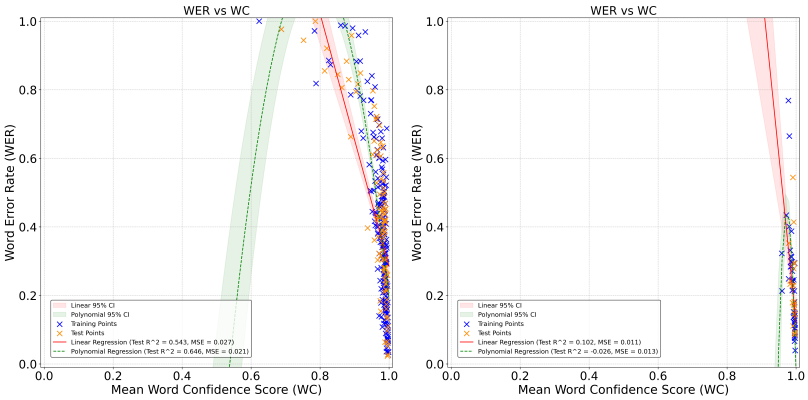

Figure 7: Relationship between WER and WC evaluated with Kraken on the OCR-D-GT (left) and VEB (right) datasets.

Dataset	r	ρ	Lin. R^2	Lin. MSE	Poly. R^2	Poly. MSE
OCR-D-GT	-0.745	-0.818	0.543	0.027	0.646	0.021
VEB	-0.385	-0.625	0.102	0.011	-0.026	0.013

Table 8: Performance metrics for the OCR-D-GT and VEB datasets in the case of Kraken.

Dataset	Lin. β_1 (p)	Lin. β_0 (p)	Poly. β_2 (p)	Poly. β_1 (p)	Poly. β_0 (p)
OCR-D-GT	-3.681 (.000)	3.966 (.000)	-20.059 (.000)	31.269 (.000)	-11.019 (.000)
VEB	-9.787 (.000)	9.907 (.000)	-678.101 (.00)	1320.854 (.00)	-642.768 (.00)

Table 9: Regression coefficients with their corresponding p-values for the OCR-D-GT and VEB datasets in the case of Kraken.

R^2 value (0.646 vs. 0.543) and a lower mean squared error, indicating that a more complex, non-linear relationship better captures the data structure. In contrast, the VEB dataset shows very low R^2 values for both models, with the polynomial regression even yielding a negative R^2 (-0.026). A negative R^2 suggests that the model fits the data worse than a simple horizontal mean line, highlighting a lack of meaningful association between the variables in this subset. The regression coefficients for the VEB dataset also display large magnitudes, further indicating instability and poor model fit. Overall, these results suggest that, unlike Calamari and Eynollah, Kraken requires a more complex model to adequately fit the OCR-D-GT data, while neither model provides a satisfactory fit for the VEB dataset.

In summary, for Eynollah, linear regression provides a reasonable fit, with stronger associations in the VEB dataset than in OCR-D-GT. Calamari achieves higher R^2 values overall, especially on OCR-D-GT, and linear models are generally sufficient. In contrast, Kraken benefits from a polynomial model for OCR-D-GT, but both models perform poorly on the VEB dataset, with negative R^2 indicating a lack of predictive power. Overall, Calamari shows the best model fit, while Kraken's results highlight the need for more complex modeling or indicate limited correlation in some cases.

E Weighted Confidence Scores Analysis

To account for the varying lengths of documents associated with each PPN and to reduce biases arising from smaller documents, we employ a weighted statistical approach throughout our evaluation. This method ensures that documents with longer texts contribute proportionally to the overall

statistics, thereby minimizing the influence of outliers and underrepresented samples.

Let $x_i \in [0, 1]$ denote the confidence score for the i-th PPN, with $w_i \geq 0$ representing its corresponding weight, defined as the number of words for word-level analysis or the number of textlines for textline-level analysis. The weighted mean confidence score is calculated as:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

To measure the dispersion of confidence scores, we compute the weighted standard deviation, where deviations are given by $d_i = x_i - \bar{x}_w$:

$$\sigma_w = \sqrt{\frac{\sum_{i=1}^{n} w_i d_i^2}{\sum_{i=1}^{n} w_i}}$$

The effective sample size, a correction accounting for the distribution of weights, is calculated as:

$$n_{\text{eff}} = \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n} w_i^2}$$

Using this, the standard error of the mean (SEM) is estimated as:

$$\text{SEM}_w = \frac{\sigma_w}{\sqrt{n_{\text{eff}}}}$$

For visualization, we construct weighted histograms by summing the weights of all confidence scores falling into each bin. The error bars for each bin are computed as:

$$\text{Error}_j = \sqrt{\sum_{i \in B_j} w_i^2}$$

This approach extends the conventional Poisson error ($\sqrt{n_j}$) used in unweighted histograms, appropriately reflecting the contribution and variability of each observation.

To further explore the distribution of confidence scores, we apply kernel density estimation (KDE) with weights. The weighted KDE at a point x is given by:

$$\hat{f}(x) = \frac{1}{h \sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \, K\left(\frac{x - x_i}{h}\right)$$

where $K(\cdot)$ is the Gaussian kernel and h is the bandwidth parameter. The resulting density curve is normalized so that its total area equals one. Additionally, we report the 25th, 50th (median), and 75th percentiles to summarize the central tendency and variability of the confidence scores.

This weighted statistical framework ensures that our analysis accurately reflects the data's heterogeneity and provides robust estimates of model performance across documents of differing sizes.

F Word and Textline Analysis

Understanding how OCR confidence varies with document size is essential for developing adaptive processing workflows, especially when working with heterogeneous and large-scale historical corpora. This analysis aims to investigate whether document length serves as a reliable predictor of OCR confidence, and to examine to what extent it reflects the underlying quality and regularity

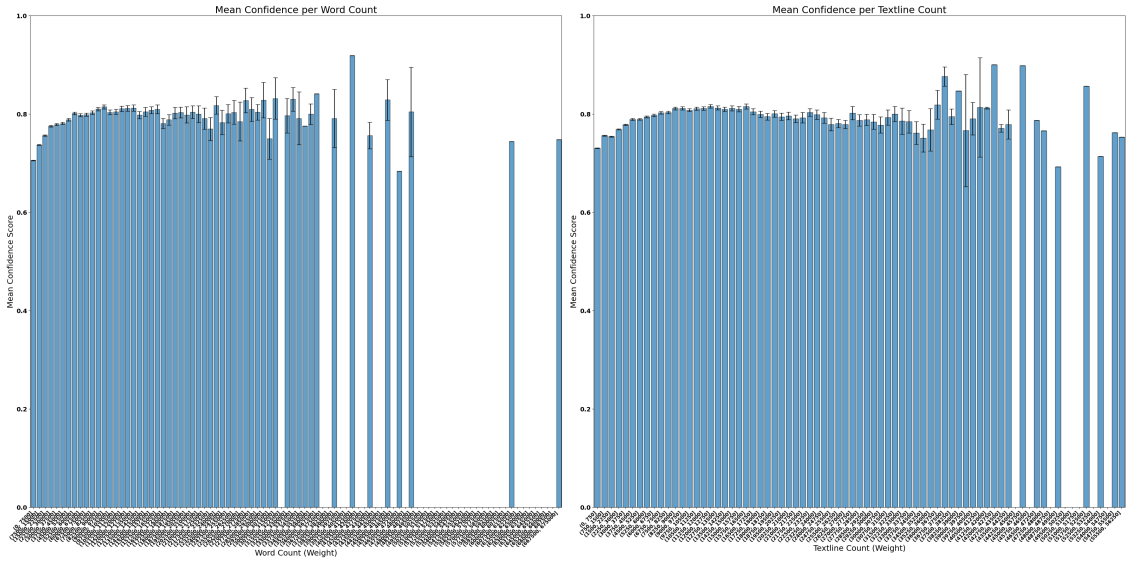

Figure 8: Relation of document length by word count (left) and textline count (right).

of the source material. Figure 8 presents two bar plots illustrating the relationship between document length, quantified by word count (left) and textline count (right), and the corresponding mean OCR confidence scores.

For each bin, the mean value of the target variable is determined, and the corresponding error bars represent the standard error of the mean.

Overall, both distributions demonstrate a consistent pattern: OCR confidence scores tend to increase with document length up to a certain point, after which the trend first stabilizes and later becomes more variable. This is especially apparent in the left plot, where mean confidence scores rise from approximately 0.70 in documents with fewer than 7,500 words to over 0.82 in medium-length documents consisting of around 52,500 to 165,000 words. Similarly, the right plot shows confidence scores improving up to about 0.81 for documents containing 6,750 to 15,750 textlines.

The narrower error bars observed in the mid-range length categories suggest that these groups are well-represented within the dataset, resulting in more reliable estimates. In contrast, greater variability and sporadic confidence scores in documents with very small or very large lengths, particularly those with extreme counts, reflect smaller sample sizes and increased uncertainty. Such patterns are common in historical corpora, where very short or very long documents often correspond to outliers such as title pages, indexes, or composite works. Further analysis reflecting the page count is presented in Appendix G.

G Page Count Analysis

To complement our analysis of word and textline counts (see Appendix F), Figure 9 examines the relationship between OCR confidence and document length, measured in terms of page count. The left panel illustrates the weighted mean confidence at word level, while the right panel depicts the corresponding confidence at the textline level. Both metrics are aggregated across binned intervals of page counts. Error bars represent the standard error of the mean, calculated based on effective sample sizes that account for the weighting scheme.

The observed patterns closely mirror those identified in the analyses of word and textline counts, yet they also provide additional insights into how document extent influences recognition quality. In both panels, there is a visible upward trend in confidence scores for documents ranging from approximately 20 to 150 pages. This increase likely reflects improved OCR performance on

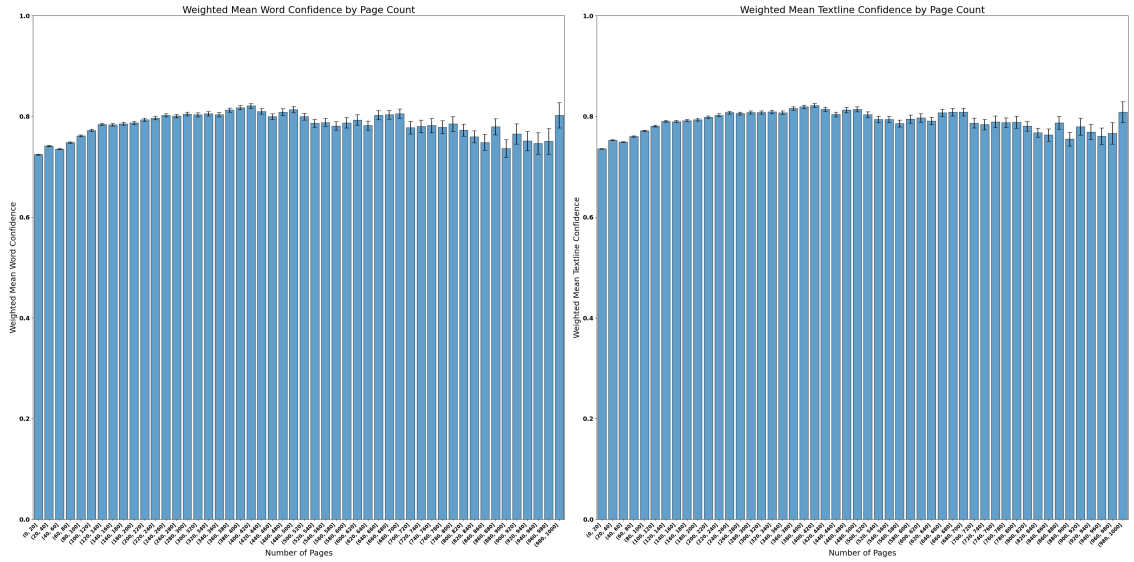

Figure 9: Weighted mean OCR confidence stratified by document page count, for word-level (left) and textline-level (right) measurements.

medium-length texts, which may benefit from more consistent typography or better preservation of original formatting.

Within the intermediate range, approximately 200 to 360 pages, confidence scores tend to plateau around 0.80–0.82, with relatively small standard errors. This stability suggests that OCR accuracy is reliably maintained within this document length segment. Conversely, documents exceeding $\sim$500 pages display more fluctuating and less stable confidence values, as evidenced by broader error bars. Such variability may stem from diverse factors, including the inclusion of ephemeral materials, title pages, or suboptimal scan quality in shorter documents, as well as composite volumes, degraded pages, or complex layouts in longer works.

An intriguing observation is the slight decline in confidence for documents exceeding approximately $\sim$700 pages, accompanied by increased uncertainty. This trend indicates that extremely long documents may pose particular challenges to OCR systems, potentially due to heterogeneity in layout, content complexity, or deterioration over extended text runs. Moreover, many of these very lengthy works are bibliographies, catalogs, dictionaries, and similar types of documents that often contain multiple fonts, languages, and complex formatting. Such characteristics can further complicate the recognition process, leading to higher error rates and reduced confidence in the OCR output for these extensive texts.

In conclusion, the analysis based on page count corroborates and extends earlier findings: medium-length documents tend to yield the highest and most consistent OCR confidence scores, whereas both shorter and longer works are associated with lower and more variable recognition quality. These insights are useful for the curation and preprocessing of large historical corpora, as they can inform strategies for selecting reliable subsets for downstream analysis or flagging documents that may require manual review or reprocessing.

H Evaluation of Genres

The dataset encompasses a total of 215 unique genres, reflecting a broad diversity of publication types. However, the distribution of publications across these genres exhibits a long-tail behavior, with a few genres dominating in publication count while many others are represented by only a small number of documents. This skewed distribution highlights the variability in genre prevalence

within the dataset.

In the German library system, genre classification for works printed before 1800 should be performed according to a list of genre terms provided by the "Arbeitsgemeinschaft Alte Drucke" (AAD);[17] these terms are part of the metadata. However, due to changing librarian practices, these genre terms are not available for all works, especially not for those printed in the 16th century. For works printed in the 19th century or later, no genre terms are available. Works without a specified genre were therefore labeled as *Unbekannt* (Unknown). To illustrate the most prominent genres, Table 10 presents the top five genres with the highest publication counts. These genres include *Unbekannt* (Unknown), *Gelegenheitsschrift* (Occasional Writing), *Leichenpredigt* (Funeral Sermon), *Lyrik* (Poetry), and *Flugschrift* (Pamphlet), which together account for a significant portion of the dataset.

Genre	Number of publications	Mean WC $\pm$ SEM	Mean TC $\pm$ SEM
Unbekannt	13137	0.804 ± 0.007	0.808 ± 0.007
Gelegenheitsschrift	6880	0.743 ± 0.009	0.751 ± 0.009
Leichenpredigt	5249	0.718 ± 0.010	0.732 ± 0.010
Lyrik	3650	0.774 ± 0.013	0.784 ± 0.013
Flugschrift	2723	0.707 ± 0.014	0.727 ± 0.014

Table 10: Top five genres with the highest publication counts in the dataset.

The *Unbekannt* category exhibits the highest average OCR confidence (0.804 word-level, 0.808 textline-level), suggesting that metadata completeness is not directly tied to recognition quality and that this category includes more recent works. *Gelegenheitsschriften* and *Leichenpredigten* show lower confidence scores (0.74–0.75 and 0.71–0.73, respectively), likely due to layout variability, typographic complexity and having little text per work. Interestingly, *Lyrik* (Poetry) exhibits higher average OCR confidence scores (0.774/0.784) than both *Gelegenheitsschrift* and *Leichenpredigt*. This suggests that the typically simpler, single-column layout and minimal typographic complexity of poetic works, along with their good preservation state, may offset potential layout-related difficulties and contribute to higher recognition confidence. *Flugschriften* (Pamphlets), although historically important, register the lowest OCR confidence among the top five genres (0.707/0.727). They are typically brief and often in degraded physical condition; combined with a wide variability in typefaces and formatting, this likely contributes to lower scores.

Across all genres, textline-level confidence scores are consistently higher than word-level scores. This may reflect the OCR engine's greater reliability in detecting interword spacing, making the segmentation and recognition of entire textlines more robust than that of individual words.

A detailed evaluation of subgenres within the dataset is provided in Appendix I.

I Evaluation of Subgenres

Figure 10 displays the weighted mean OCR confidence scores for various genre-subgenre combinations, with word-level (blue) and textline-level (red) scores shown alongside standard error bars. The analysis covers the five genres in the dataset that include subgenre annotations: *Gelegenheitsschrift* (Occasional Writing), *Streitschrift* (Polemic Writing), *Kommentar* (Commentary), *Ordensliteratur* (Publications by Religious Orders), and *Dissertation* (Dissertation). These account for a subset of the 215 total genres in the corpus, indicating that subgenre-level classification is comparatively rare.

[17] `https://verbundwiki.gbv.de/spaces/GAD/pages/73990159/Gattungsbegriffe+der+Arbeitsgemeinschaft+Alte+Drucke+beim+GBV+und+SWB`

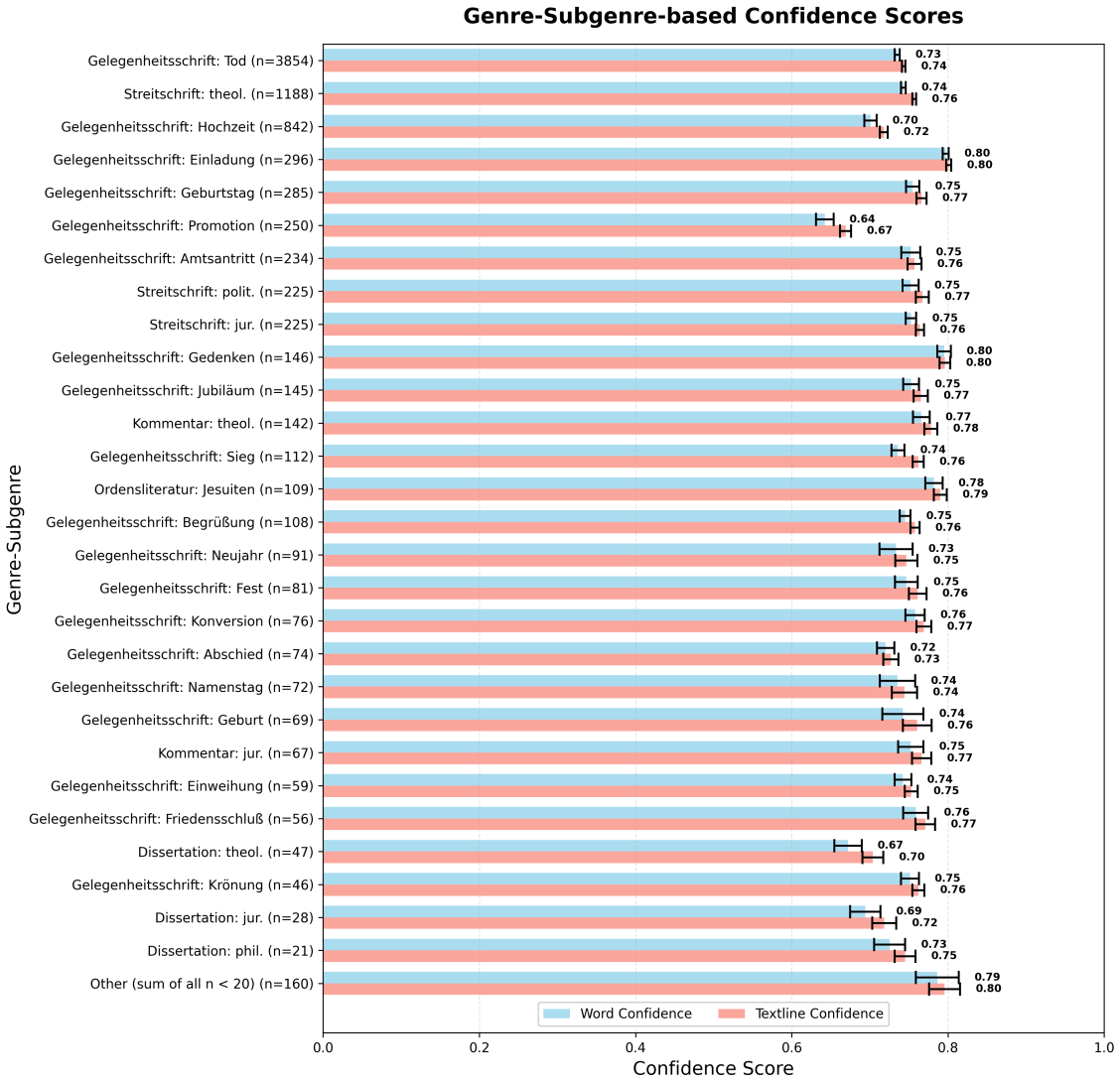

Figure 10: Evaluation of genre-subgenre combinations.

The results highlight substantial variability in OCR performance across subgenres. High confidence scores, exceeding 0.80, are observed for subgenres such as *Gelegenheitsschrift: Einladung* (Invitation), *Gelegenheitsschrift: Gedenken* (Memorial), and *Kommentar: theol.* (Theological Commentary), likely reflecting the presence of well-structured and typographically consistent content. In contrast, lower confidence values, particularly below 0.70, are found in subgenres such as *Gelegenheitsschrift: Promotion* (Doctoral Promotion) and *Dissertation: theol.* (Theological Dissertation), which may contain more complex or deteriorated source material. A consistent pattern emerges across nearly all subgenres: textline-level confidence scores are slightly but systematically higher than their word-level counterparts.

The aggregated ”Other” category, summarizing subgenres with fewer than 20 instances each, shows relatively high confidence scores ($\sim$0.79 for words, $\sim$0.80 for textlines), although its interpretive value is limited due to the heterogeneity of the underlying documents.

Overall, these findings suggest that subgenre-aware approaches may enhance quality control in large-scale digitization efforts.

J Evaluation of the 17th-Century Data

Analysis of confidence score distributions across 11,993 PPNs in the 17th century (see Figure 11) reveals consistently high recognition confidence for both words and textlines, with mean scores predominantly ranging from 0.6 to 0.9. The weighted mean confidence score is 0.706 ± 0.006 for words and 0.726 ± 0.007 for textlines, both slightly lower than the overall dataset means of 0.785 $\pm$ 0.004 and 0.791 $\pm$ 0.004, respectively. This suggests that the subset contains more complex or challenging material. The low standard deviations at both levels indicate stable and reliable confidence estimates, with generally low variability across instances.

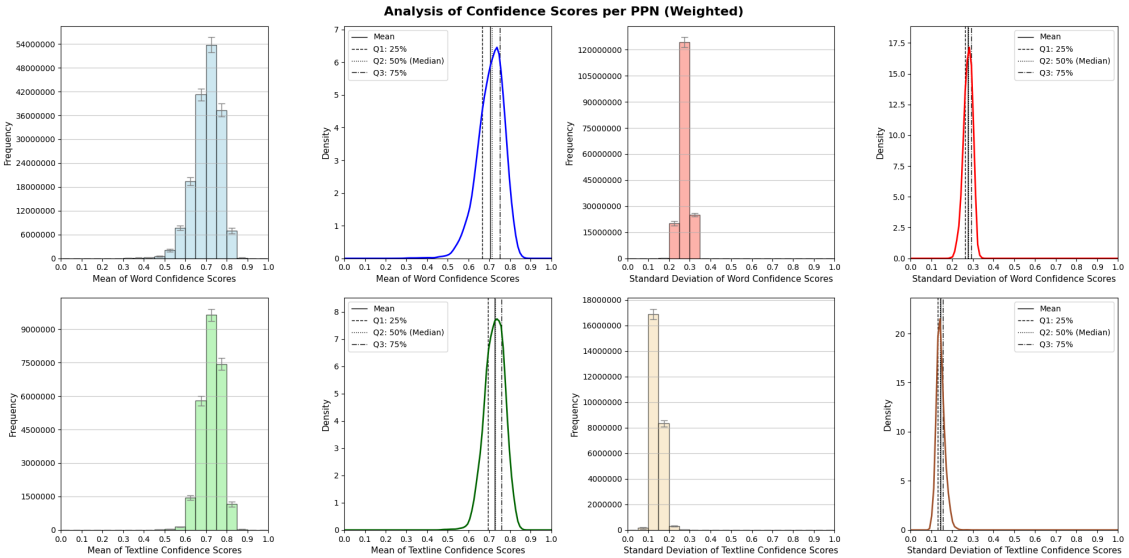

Figure 11: Evaluation of the historical dataset covering the years 1601–1700.

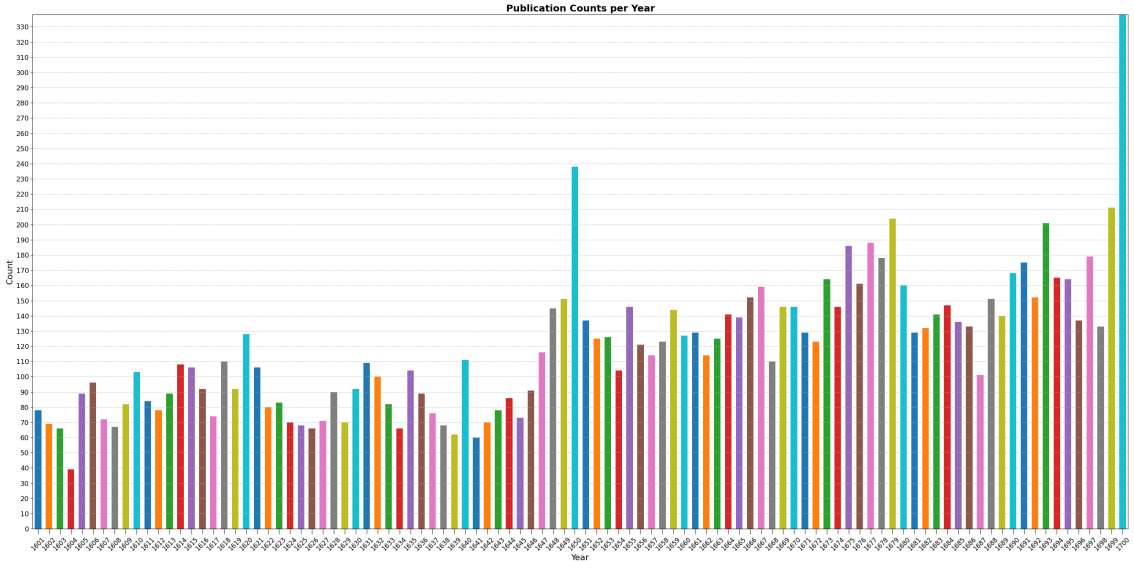

Figure 12: Publication counts covering the years 1601–1700.

Figure 12 displays publication counts per year throughout the 17th century, illustrating a general upward trend in publication activity. However, a closer examination of the period corresponding to the Thirty Years' War (1618–1648) reveals a noticeable stagnation and even a slight decline

in publication counts compared to the preceding and subsequent years. This pattern suggests a potential disruptive effect of the war on the production and dissemination of printed materials, likely due to widespread social, economic, and political instability.

Following the end of the war, publication activity gradually resumes its upward trajectory, culminating in a pronounced peak at the end of the century. Notably, this post-war period coincides with the beginning of the New High German era (starting around 1650), which marked a significant phase in the standardization of both written and spoken German. The increasing linguistic uniformity, alongside growing stability and the recovery of intellectual and cultural life, likely contributed to the revitalization of the print industry. This convergence of sociopolitical recovery and linguistic development may explain the sustained rise in publication output during the second half of the century.

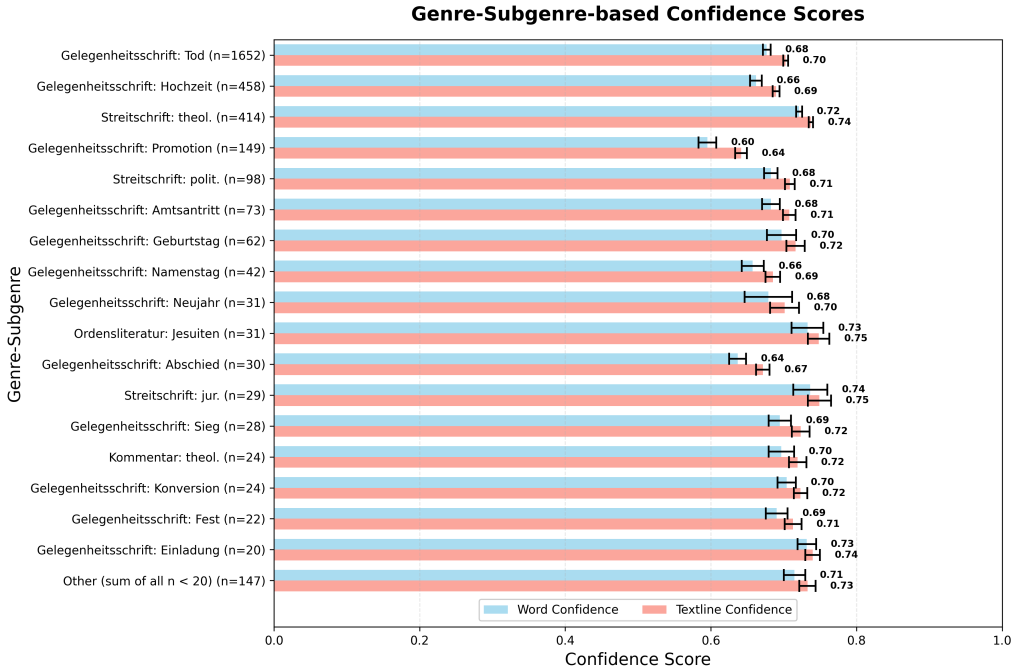

Figure 13: Evaluation of genre-subgenre combinations covering the years 1601–1700.

Figure 13 shows that confidence scores vary notably across subgenres, with textline confidence consistently exceeding word confidence. Subgenres such as *"Gelegenheitsschrift: Tod"* and *"Gelegenheitsschrift: Hochzeit"* achieve the highest scores, likely due to their standardized layouts and clearer print quality. In contrast, subgenres with lower confidence, such as *"Gelegenheitsschrift: Abschied"*, may reflect greater variability in formatting or print quality. Compared to the whole dataset, the 17th-century subgenres generally exhibit lower confidence scores, with fewer subgenres exceeding 0.80, indicating that OCR performance is more challenged by the material from this period.

As shown in Figure 14, the 17th century exhibits a distinctive pattern in genre-based OCR confidence scores. Out of the 215 genres present in the overall dataset, 121 are represented in this century, reflecting a remarkable diversity of printed material during this period. Unlike other centuries, where *"Unbekannt"* is typically the most prevalent genre and achieves the highest confidence scores, *"Leichenpredigt"* emerges as the leading genre in both frequency and recognition certainty in the 17th century. Specifically, *"Leichenpredigt"* achieves mean confidence scores of approximately 0.70 for both word and textline recognition. This may be attributed to the relatively standardized structure commonly found in funeral sermons of the era, which facilitate more ac-

curate text recognition. In contrast, *"Einblattdruck"* shows much lower mean confidence scores (0.59 for words, 0.63 for textlines), suggesting that single-sheet prints are more challenging for OCR due to their varied layouts, typefaces used, print quality, and decorative elements.

If compared to the results for the entire dataset (Table 10), it becomes evident that the mean confidence scores for the most frequent genres in the 17th century are generally lower than those observed across all centuries. For example, *"Unbekannt"* and *"Leichenpredigt"* have mean word confidence scores of 0.804 and 0.718, respectively, in the full dataset, compared to 0.71 and 0.70 in the 17th century.

K Discussion of the Tesseract models

For the bulk of the processed works discussed in this paper, we used Tesseract 4 with the `deu_frak` model and to a lesser extent the `german_print` model. The `deu_frak` model is a newer LSTM-based Tesseract 4 model trained from scratch on actual Ground Truth, whereas `deu_frak` is a conversion of the Tesseract 3 model trained on synthetic data. These models are not directly comparable in operation, including confidence calculation. In hindsight, the choice of `deu_frak` was suboptimal but we finished the processing for consistency and switched to `german_print` for newer datasets. While the confidence output is still consistent within works processed with either model, considering the opaque training parameters and legacy codebase used for ported Tesseract 3 models, we strongly recommend against directly comparing Tesseract 3 and Tesseract 4 model confidence behavior.

Beyond the historical printing developments described in Appendix 5.2, the observed trend of increase in OCR accuracy is also influenced by the training data of the OCR model. The `deu_frak` model, synthetically trained on 19th and 20th century Fraktur fonts, performs better on more recent printed material whose visual features align more closely with the model's learned representations. This correspondence further reinforces the improvement in recognition quality over time.

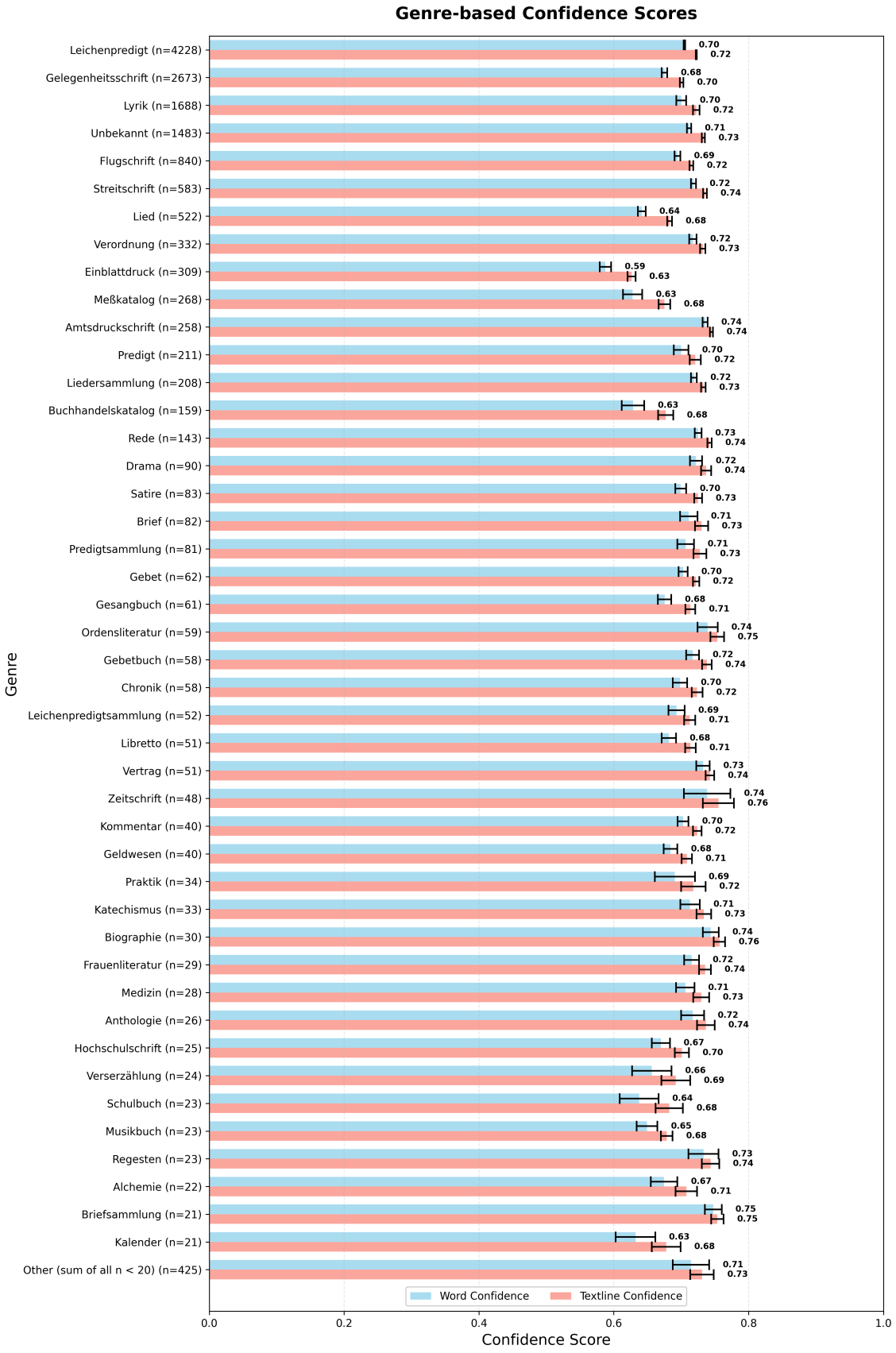

Figure 14: Evaluation of the genres covering the years 1601–1700.