“Ein Freinttlichen Trunck”: the Zechzettel dataset for handwritten text recognition and information retrieval

Fabio Mariani, Helmut Graser, B. Ann Tlusty, Regina Dauser, Annemarie Friedrich

Angaben zur Veröffentlichung / Publication details:

Mariani, Fabio, Helmut Graser, B. Ann Tlusty, Regina Dauser, and Annemarie Friedrich. 2026. “‘Ein Freinttlichen Trunck’: the Zechzettel dataset for handwritten text recognition and information retrieval.” In *DHd 2026: Nicht nur Text, nicht nur Daten - 12. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum e.V., Universität Wien, Wien, Österreich, 23.–27. Februar 2026*, edited by Silke Schwandt, Gabriel Viehhauser, Tara Andrews, and Thomas J. J. Wallnig, 493–94. Geneva: Zenodo. https://doi.org/10.5281/zenodo.18702907.

"Ein Freinttlichen Trunck": The Zechzettel Dataset for Handwritten Text Recognition and Information Retrieval

Mariani, Fabio
fabio.mariani@uni-a.de
University of Augsburg, Deutschland
ORCID: 0000-0002-7382-0187

Graser, Helmut
helmut.graser@philhist.uni-augsburg.de
University of Augsburg, Deutschland
ORCID: 0009-0001-4607-342X

Tlusty, B. Ann
tlusty@bucknell.edu
Bucknell University, Pennsylvania, Vereinigte Staaten von Amerika
ORCID: 0000-0002-1193-7558

Dauser, Regina
regina.dauser@uni-a.de
University of Augsburg, Deutschland
ORCID: 0000-0002-3289-4964

Friedrich, Annemarie
annemarie.friedrich@uni-a.de
University of Augsburg, Deutschland
ORCID: 0000-0001-8771-7634

Introduction

Handwritten historical documents exhibit visual and linguistic features that vary by time, place, and author, making it difficult to develop models that generalize across such variation. Despite recent advances in Handwritten Text Recognition (HTR), evaluation still relies largely on Character Error Rate (CER) and Word Error Rate (WER) over held-out test data. While useful for assessing transcription accuracy, these metrics offer limited insight into how well models generalize across the heterogeneous visual and linguistic dimensions that characterize handwritten texts (Garrido-Munoz and Calvo-Zaragoza, 2025; Hodel et al., 2021).

Evaluating HTR systems in such a context requires resources that reflect the real-world variability of historical manuscripts and support analysis beyond transcription accuracy. As part of this investigation, we are developing the *Zechzettel* dataset, based on a historical collection of approximately 1,800 handwritten petitions produced in Augsburg between 1583 and 1605 and preserved at the Stadtarchiv Augsburg. These petitions requested permission for drinking outside the city limits, a practice regulated and taxed by local authorities.

The collection captures a wide range of handwriting styles and linguistic variants (Figure 1). While some petitions in the *Zechzettel* collection were written by professional scribes, many reflect the handwriting of ordinary citizens, using non-standard orthography and local dialects, exemplifying forms of script from below. Originally assembled for historical and linguistic research, the dataset includes high-resolution manuscript images, expert transcriptions, and qualitative annotations supporting the study of interactions between visual and linguistic features in HTR and downstream tasks such as information retrieval.

Different examples from the *Zechzettel* collection. Original manuscripts held at the Stadtarchiv Augsburg.

Background

Evaluating the quality of HTR systems requires moving beyond standard metrics like CER and WER, to account for additional dimensions such as linguistic plausibility, visual input quality, and performance in downstream tasks. However, these dimensions are typically examined in isolation, limiting our understanding of how they interact in complex historical documents.

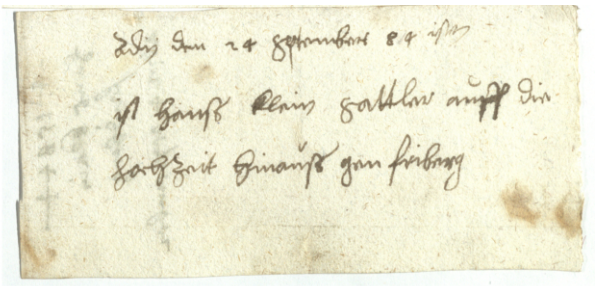
Linguistically motivated approaches use lexicon-based and language model-based metrics to evaluate the plausibility of HTR outputs beyond standard error rates (Ströbel et al., 2022). These include token ratios, character n-gram frequencies, and language models (pseudo-)perplexity scores, serving as proxies for transcription plausibility.

Visually grounded approaches, by contrast, assess document image quality based on features that affect readability and recognition. Document Image Quality Assessment (DIQA) aims to estimate legibility without relying on transcriptions, extracting quality signals such as blur, noise, or contrast that correlate with HTR performance (Alaei et al., 2023).

Finally, evaluations based on information retrieval assess HTR quality through the reliability of downstream tasks like named entity recognition (NER). Notably, even when overall CER or WER scores are similar across different types of visual degradation, the impact on NER performance can vary significantly (Hamdi et al., 2020). This underscores that transcription quality cannot be fully captured by error rates alone; instead, it reflects the complex interaction between visual distortions and linguistic structures that ultimately shape information retrieval outcomes.

Dataset Preparation

The *Zechzettel* collection was compiled and transcribed by an expert in historical German linguistics and annotated with corresponding metadata. [1] Figure 2 illustrates a typical example, showing a petition from 1584 by Hanß Klein, described in the original annotation as having a "very scrawled hand" and "gray ink." Only the front side is available as a digital image, while the reverse exists in transcription only. In this illustration, image and text are manually aligned to convey the type of material included in the dataset; this alignment, however, is not present in the original collection.



Notes: Sehr krakelige Hand, graue Tinte; eigenhändig.
Transcription:
Ady den 24 September 84 Jar [?]
ist Hanß klein Sattler auff die
hochzeit Hinauß gen friberg
Reverse:
+ 1584 +
hans klain *vergonst*
fridberg
adj 24 september

Example from the *Zechzettel* collection, petition by Hanß Klein, 1584. Original manuscript held at the Stadtarchiv Augsburg.

To transform the *Zechzettel* collection into an operable dataset for HTR, we employ a semi-automatic processing pipeline. Because individual scans frequently contain multiple documents, the first step consists of extracting documents from each scan. Each document image is then segmented into text lines to obtain line-level spatial coordinates. These coordinates enable alignment between document images and expert transcriptions, with editorial uncertainties preserved as structured metadata. In addition, d ocument quality is encoded through a structured taxonomy capturing features such as handwriting style, ink characteristics, and writing tools.

The *Zechzettel* dataset is currently under development and provides ground truth for investigating HTR performance under real-world historical variation. By combining visual features and transcriptions, it supports analyses beyond standard error rates and enables future research on HTR evaluation and historical information retrieval.

Footnotes

1. The transcription and annotation work were carried out by Dr. Helmut Graser as part of an interdisciplinary project in collaboration with Prof. Dr. B. Ann Tlusty.

Bibliography

Alaei, Alireza, Vinh Bui, David Doermann, and Umapada Pal . 2023. "Document Image Quality Assessment: A Survey." ACM Computing Surveys , 56(2): 1–36.

Garrido-Munoz, Carlos, and Jorge Calvo-Zaragoza. 2025. "On the Generalization of Handwritten Text Recognition Models." In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) , 15275–15286.

Hamdi, Ahmed, Axel Jean-Caurant, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. 2020. "Assessing and Minimizing the Impact of OCR Quality on Named Entity Recognition." In Digital Libraries for Open Knowledge: 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25–27, 2020. Proceedings , 87–101.

Hodel, Tobias, David Schoch, Christa Schneider, and Jake Purcell. 2021. "General Models for Handwritten Text Recognition: Feasibility and State-of-the-Art. German Kurrent as an Example." Journal of Open Humanities Data , 7: 1–10.

Ströbel, Phillip Benjamin, Martin Volk, Simon Clematide, Raphael Schwitter, Tobias Hodel, and David Schoch. 2022. "Evaluation of HTR Models Without Ground Truth Material." In Proceedings of the Thirteenth Language Resources and Evaluation Conference , 4395–4404.