JobAds Gold Standard Dataset

The Gold Standard is a manually corrected transcription dataset of 14,873 historical newspaper job-advertisement segments created within the JobAds project at the University of Graz, with the support of student assistants between 2024 and 2025. It was developed to provide a high-quality textual reference for the evaluation of Optical Character Recognition (OCR) models and for subsequent text-mining experiments on historical job advertisements.

The Gold Standard is based on the Ground Truth dataset for page segmentation created within the JobAds project. The Ground Truth contains 14,985 manually annotated segments representing individual job advertisements identified on historical newspaper pages. The Ground Truth was published as part of the CHR2024 conference paper by Venglarova, Adam, Mölzer et al. (2024) and is publicly available in CSV format via the https://github.com/JobAds-FWFProject/Ground-Truth-CHR2024 repository on GitHub. The dataset can be reused under the CC BY 4.0 license.

Whereas the Ground Truth dataset identifies and categorizes the location of advertisements on newspaper pages, the Gold Standard provides a corrected transcription of the text contained within these annotated segments. The two datasets are therefore complementary: the Ground Truth specifies where an advertisement is located and what type of advertisement it is (job offer, job search, service offer, heading, mediation, indicator), while the Gold Standard specifies what the advertisement text contains.

The raw OCR transcriptions used as the starting point for the Gold Standard were generated with the frak2021 OCR model (Mannheim University Library, 2021). These OCR outputs were manually checked and corrected using the Transkribus platform (Kahle et al., 2017). The transcription guidelines aimed to reproduce the printed original as faithfully as possible rather than to normalize the text according to modern spelling or typographical conventions.

Accordingly, historical and typographical characteristics of the source material were preserved. For example, the historical long *s* (*ſ*) was retained rather than converted to the modern *s*. Original hyphenation, spacing, spelling, and printing errors were likewise preserved. This approach ensures that the Gold Standard remains as close as possible to the historical source and provides an appropriate reference for evaluating OCR systems on historical documents. In addition to correcting OCR errors, annotators corrected the reading order of individual text lines where this had not been correctly recognized by the OCR system. Completely unreadable words were represented by `####`.

As with the Ground Truth, the Gold Standard was created in a single annotation/correction round and was not subsequently subjected to systematic verification due to limited human resources. Consequently, occasional errors in the manual corrections or overlooked OCR errors may remain in the dataset. The extent of such errors cannot currently be quantified. The Gold Standard should therefore be understood as a carefully manually corrected reference dataset rather than as an absolutely error-free transcription.

Data Format

The Gold Standard is provided in CSV format. Each row corresponds to one of the 14,873 annotated job-advertisement segments and contains metadata describing the source newspaper and page, the location and category of the advertisement, the original OCR output, and the manually corrected transcription. This is slightly fewer than the 14,985 segments in the Ground Truth, as a small number of segments were illegible in the source material and could not be transcribed, and were therefore discarded from the Gold Standard.

The dataset contains the following columns:

Column	Description
`pub`	Identifier of the newspaper publication (e.g., `gtb` for *Grazer Tagblatt*).
`pub_date`	Publication date of the newspaper (e.g., 02031860 for 2.3.1860).
`page_num`	Page number of the newspaper page.
`page_id`	Unique identifier of the newspaper page, combining the publication, date, and page information.
`uuid`	Unique identifier of the individual advertisement segment.
`label`	Category assigned to the advertisement in the Ground Truth: job offer, job search, service offer, heading, or mediation.
`x`	Horizontal coordinate of the advertisement's bounding box on the page, measured from the top-left corner.
`y`	Vertical coordinate of the advertisement's bounding box on the page, measured from the top-left corner.
`width`	Width of the advertisement's bounding box.
`height`	Height of the advertisement's bounding box.
`ocr`	Raw OCR transcription generated by the frak2021 OCR model, including line breaks.
`ocr_without_lb`	Raw OCR transcription with line breaks removed.

`text`	Manually corrected Gold Standard transcription, preserving line breaks and the characteristics of the printed original.
`text_without_lb`	Manually corrected Gold Standard transcription with line breaks removed.
`iiif`	IIIF link leading to the image of the corresponding advertisement, allowing the transcription to be compared directly with the original newspaper source.

The coordinate fields (`x`, `y`, `width`, and `height`) describe the rectangular bounding box of each advertisement segment on the corresponding newspaper page. The `ocr` and `text` fields preserve line breaks, while the corresponding `_without_lb` fields provide versions without line breaks, which can be useful for automated text processing and OCR evaluation.

The `ocr` and `text` fields allow direct comparison between the automatically generated OCR output and the manually corrected Gold Standard transcription. The `iiif` field provides direct access to the corresponding advertisement image, supporting manual verification and further research using the original source material.

References

Kahle, P. et al. (2017). Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). pp. 19–24. 10.1109/ICDAR.2017.307.

Mannheim University Library. (2021). Frak2021.https://ub-backup.bib.uni-mannheim.de/~stweil/tesstrain/frak2021/tessdata_best/frak2021-0.905.traineddata (accessed 20 November 2023).

Venglarova, K., Adam, R., Mölzer, W., et al. (2024). Who Advertises in Newspapers? Data Criticism in Mining Historical Job Ads. In *Proceedings of the Computational Humanities Research Conference 2024*. Computational Humanities Research 2024. Aarhus, Denmark: CEUR-WS, pp. 788–801.ceur-ws.org/Vol-3834/.