JobAds: From Digitized Newspapers to Economic Analysis

Venglarova, Klara

klara.venglarova@uni-graz.at
Universität Graz, Österreich
ORCID: 0009-0007-6441-7795

Adam, Raven

raven.adam@uni-graz.at
Universität Graz, Österreich
ORCID: 0000-0001-7841-2601

Mölzer, Wiltrud

wiltrud.moelzer@uni-graz.at
Universität Graz, Österreich
ORCID: 0009-0002-9517-4531

Balasubramanian, Saranya

saranya.balasubramanian@uni-graz.at
Universität Graz, Österreich

Füllsack, Manfred

manfred.fuellsack@uni-graz.at
Universität Graz, Österreich
ORCID: 0000-0002-7772-4061

Kleinert, Jörn

joern.kleinert@uni-graz.at
Universität Graz, Österreich
ORCID: 0000-0002-1167-9245

Vogeler, Georg

georg.vogeler@uni-graz.at
Universität Graz, Österreich
ORCID: 0000-0002-1726-1712

The JobAds Project (FWF P35783) explores the development of the Austrian labor market between 1850-1950 through extracting and analysing job advertisements from digitized newspapers from this period. Our research consists of two phases: first, converting data into machine-readable text, and second, extracting job advertisements and analyzing them. With the first phase nearing completion, we are now entering into the second phase.

Several past or present projects, such as *NewsEye* (Doucet et al. 2020), *Impresso* (Ehrmann et al. 2020) or *Historical Ink* (Manrique-Gomez et al. 2024), focus on digitized newspapers, aiming to create large collections or develop tools for automated processing. However, challenges in layout analysis and optical character recognition (OCR) for historical newspapers still remain far from being perfectly resolved (Torget 2023; Wevers 2023). This also proved to be a significant constraint in our efforts of extracting job advertisements.

Our process began with selecting images containing job advertisements from the ANNO corpus (Österreichische Nationalbibliothek 2021). These images were segmented and converted into machine-readable text, which was refined using a post-correction model. Each step required applying or fine-tuning existing models and evaluating their performance for our data and use-case.

Initially, we manually reviewed sampled issues from 29 newspaper titles across the defined time-span to identify pages containing job advertisements. Observing patterns in the appearance of job-ad sections, we selected pages from the entire corpus based on these patterns, aiming for a high recall. Later, we fine-tuned a visual transformers-based model (Lewis et al. 2006), according to which about 34% of the pre-selected pages contained job advertisements, resulting in about 1,36 million pages relevant to our research.

Layout analysis proved to be a critical step, as job-ads pages often have irregular layouts due to space-saving practices, decorative frames, or rotated content. Therefore, segmentation algorithms tended to separate headings from ad text, or to merge multiple ads. To quantify this problem, we manually annotated almost 15,000 job advertisements across all our newspapers and years. This ground-truth dataset was published in the CHR2024 conference paper (Venglarova, Adam, Mölzer, et al. 2024).

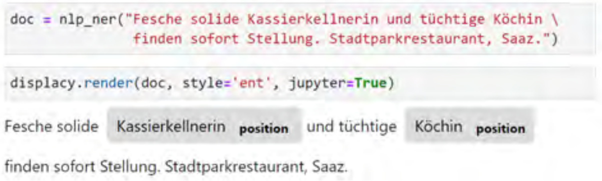
To evaluate segmentation, we developed a methodology that examines text presence in non-intersecting areas of predicted regions and their ground truth, rather than focusing solely on their intersection. Using the model described in (Venglarova, Adam, Balasubramanian, et al. 2024), we compared segmentation results from the Eynollah model (Rezanezhad et al. 2023), the default ANNO corpus segments (Österreichische Nationalbibliothek 2021), and Tesseract (Smith 2009). The comparison was based on a random sample of 250 manually annotated segments. The Eynollah model achieved the highest accuracy (Tab. 1) and we adopt this model for further work. However, the solutions for layout analysis tailored for historical newspapers are quickly evolving at the time of writing this article, and we expect even better results in the near future (Sun et al. 2024; Girdhar, Coustaty, and Doucet 2023).

Table 1: Comparison of segmentation models' performance. Tested on a sample of 250 manually annotated segments.

Segments from	Accuracy
ANNO Corpus	45.312%
Eynollah	72.074%
Tesseract	20.518%

Next, we converted segmented images into machine-readable text using OCR with the frak2021 model (Mannheim University Library 2021) and applied post-correction using a hmbyt5-preliminary model fine-tuned on the ID-CAR2019-POCR dataset for OCR correction (Rigaud et al. 2019). To evaluate the quality of the OCRed text and the OCR correction output, we created a gold standard dataset by manually correcting the text in the ads, using the Transkribus platform (Kahle et al. 2017). This provides us with insights into common error patterns in the OCR stage, such as confusing letters *n* and *u* or *f* and *ſ*, which helped us improve the quality of the text in the post-correction stage.

In the analytical phase, our first task was to extract job titles from the advertisements. We compared four methods: a dictionary-based approach, linguistic structure analysis using parts-of-speech and dependency parsing, and machine-learning (ML) approaches, namely named entity recognition (NER) (Fig. 1) and text generation. Achieving similarly high accuracies with both ML approaches (Venglarova, Adam, and Vogeler 2024), this step enables us to observe changing trends of advertising certain jobs by relative frequencies.



```
doc = nlp_ner("Fesche solide Kassierkellnerin und tüchtige Köchin \
                finden sofort Stellung. Stadtparkrestaurant, Saaz.")
```

```
displacy.render(doc, style='ent', jupyter=True)
```

```
Fesche solide  Kassierkellnerin position  und tüchtige  Köchin position
finden sofort Stellung. Stadtparkrestaurant, Saaz.
```

Fig. 1: Extracting position names with the NER approach.

Next steps in our research include the following:

1. Complete the gold standard and evaluate the OCR across years and publications.
2. Explore effects of erroneous segmentation and options for its improvement.
3. Compare post-correction techniques, such as dictionary-based, rule-based, and BERT-based methods.
4. Train a model to classify text segments as job-ad or non-job-ad regions.
5. Extract and analyze requirements related to different positions and their evolution over time.
6. Apply BERT topic modeling to gain insights into main topics across the corpus.
7. Explore the role of abbreviations and their effects on NLP tasks.

These steps will allow us to compare changing trends in the labor market over time, focusing on job requirements and qualifications, regional and temporal differences, and gender inequalities. Alongside performing and evaluating these technical steps, we are gathering information about the historical and economic context of job advertisements in newspapers. Matches between vacancies and job seekers were realized through several channels (Wadauer, Buchner, and Mejstrik 2012), with only about 30% arranged through newspaper advertisements (Mölzer and Kleinert 2024). While newspapers dominated for white-collar job searches (Faust 1986), blue-collar workers were underrepresented in this channel. Thus, our findings will reflect only a subset of the labor market. Other factors such as the role of private agencies, the political orientation of newspapers, their geographical reach, temporal coverage, social focus, and intended readers remain important topics for further investigation.

Bibliographie

Doucet, Antoine, Martin Gasteiner, Mark Granroth-Wilding, Max Kaiser, Minna Kaukonen, Roger Labahn, Jean-Philippe Moreux, et al. 2020. ‘NewsEye: A Digital Investigator for Historical Newspapers’. In . Alliance of Digital Humanities Organizations (ADHO). https://doi.org/10.5281/zenodo.3895269.

Ehrmann, Maud, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel, and Raphaël Barman. 2020. ‘Language Resources for Historical Newspapers: The Impresso Collection’. In *Proceedings of the 12th Language Resources and Evaluation Conference* , 958–68. European Language Resources Association (ELRA). https://doi.org/10.5167/uzh-191270.

Faust, Anselm. 1986. ‘Arbeitsmarktpolitik Im Deutschen Kaiserreich: Arbeitsvermittlung, Arbeitsbeschaffung Und Arbeitslosenunterstützung 1890-1918’. *(No Title)* .

Girdhar, Nancy, Mickaël Coustaty, and Antoine Doucet. 2023. ‘STRAS: A Semantic Textual-Cues Leveraged Rule-Based Approach for Article Separation in Historical Newspapers’. In *Leveraging Generative Intelligence in Digital Libraries: Towards Human-Machine Collaboration* , edited by Dion H. Goh, Shu-Jiun Chen, and Suppawong Tuarob, 89–105. Singapore: Springer Nature Singapore.

Kahle, P., S. Colluto, G. Hackl, and G. Mühlberger. 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)* , 04:19–24. https://doi.org/10.1109/ICDAR.2017.307.

Lewis, David D., Gady Agam, Shlomo Engelson Argamon, Ophir Frieder, David A. Grossman, and Jefferson Heard. 2006. ‘Building a Test Collection for Complex Document Information Processing’. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* .

Mannheim University Library. 2021. ‘Frak2021’. https://ub-backup.bib.uni-mannheim.de/~stweil/tesstrain/frak2021/tessdata_best/frak2021-0.905.traineddata.

Manrique-Gomez, Laura, Tony Montes, Arturo Rodriguez Herrera, and Ruben Manrique. 2024. ‘Historical Ink: 19th Century Latin American Spanish Newspaper Corpus with LLM OCR Correction’. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities* , edited by Mika Hämäläinen, Emily Öhman, So Miyagawa, Khalid Alnajjar, and Yuri Bizzoni, 132–39. Miami, USA: Association for Computational Linguistics. https://aclanthology.org/2024.nlp4dh-1.13.

Mölzer, Wiltrud, and Jörn Kleinert. 2024. 'Emergence of the Austrian Labor Market'. https://static.uni-graz.at/fileadmin/_files/_project_sites/_historical-job-ads/Emergence_Austrian_labor_market.pdf.

Österreichische Nationalbibliothek. 2021. 'ANNO Historische Zeitungen Und Zeitschriften'. 2021. https://anno.onb.ac.at/.

Rezanezhad, Vahid, Konstantin Baierer, Mike Gerber, Kai Labusch, and Clemens Neudecker. 2023. 'Document Layout Analysis with Deep Learning and Heuristics'. In *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing HIP 2023, San José, US, August 25-26, 2023, ACM.* https://doi.org/10.1145/3604951.3605513.

Rigaud, Christophe, Antoine Doucet, Mickael Coustaty, and Jean-Philippe Moreux. 2019. 'ICDAR 2019 Competition on Post-OCR Text Correction'. In *Proceedings of the 15th International Conference on Document Analysis and Recognition (2019)* .

Smith, Ray. 2009. 'Hybrid Page Layout Analysis via Tab-Stop Detection'. In *ICDAR '09: Proceedings of the 2009 10th International Conference on Document Analysis and Recognition* , 241–45. Washington, DC, USA: IEEE Computer Society. http://dx.doi.org/10.1109/ICDAR.2009.257.

Sun, Wenjun, Hanh Thi Hong Tran, Carlos-Emiliano González-Gallardo, Mickaël Coustaty, and Antoine Doucet. 2024. 'LIAS: Layout Information-Based Article Separation in Historical Newspapers'. In *The 28th International Conference on Theory and Practice of Digital Libraries* , 15177:256–72. Lecture Notes in Computer Science. LJUBLJANA, Slovenia: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72437-4_15.

Torget, Andrew. 2023. 'Mapping Texts: Examining the Effects of OCR Noise on Historical Newspaper Collections'. In *Digitised Newspapers – A New Eldorado for Historians?* , edited by Estelle Bunout, Maud Ehrmann, and Frédéric Clavert, 47–66. https://doi.org/10.1515/9783110729214-003.

Venglarova, Klara, Raven Adam, Saranya Balasubramanian, and Georg Vogeler. 2024. 'Quantifying Page Segmentation Quality in Historical Job Advertisements Retrieval'. https://inria.hal.science/hal-04560463.

Venglarova, Klara, Raven Adam, Wiltrud Mölzer, Saranya Balasubramanian, Jörn Kleinert, Manfred Füllsack, and Georg Vogeler. 2024. 'Who Advertises in Newspapers? Data Criticism in Mining Historical Job Ads'. In *Proceedings of the Computational Humanities Research Conference 2024* , 788–801. Aarhus, Denmark: CEUR-WS. ceur-ws.org/Vol-3834/.

Venglarova, Klara, Raven Adam, and Georg Vogeler. 2024. 'Extracting Position Titles from Unstructured Historical Job Advertisements'. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities* , edited by Mika Hämäläinen, Emily Öhman, So Miyagawa, Khalid Alnajjar, and Yuri Bizzoni, 75–84. Miami, USA: Association for Computational Linguistics. https://aclanthology.org/2024.nlp4dh-1.8.

Wadauer, Sigrid, Thomas Buchner, and Alexander Mejstrik. 2012. 'The Making of Public Labour Intermediation: Job Search, Job Placement, and the State in Europe, 1880–1940'. *International Review of Social History* 57 (S20): 161–89. https://doi.org/10.1017/S002085901200048X.

Wevers, Melvin. 2023. 'Mining Historical Advertisements in Digitised Newspapers'. In *Digitised Newspapers – A New Eldorado for Historians?* , edited by Estelle Bunout, Maud Ehrmann, and Frédéric Clavert, 227–52. https://doi.org/10.1515/9783110729214-011.