Who Advertises in Newspapers? Data Criticism in Mining Historical Job Ads

Klara Venglarova^{1,*}, Raven Adam^{1,*}, Wiltrud Mölzer^{1,*}, Saranya Balasubramanian¹, Jörn Kleinert¹, Manfred Füllsack¹ and Georg Vogeler¹

¹University of Graz, Austria

Abstract

Digitized newspapers are a source of unique and rich historical data but pose significant challenges in the interpretation of results obtained through their mining. The JobAds project (FWF P35783) explores the evolution of the labor market through job advertisements from digitized newspapers between 1850-1950, aiming to reveal regional and temporal trends in job offers, required skills, media strategies, and social aspects such as gender-specific ads. Using the ANNO corpus, we selected 29 newspapers with the most editions. Their processing involved job ads pages preselection, layout segmentation, optical character recognition (OCR), and post-correction, each introducing potential biases due to varying efficiency of these processes. Additionally, the inherent bias of newspapers as historical sources must be considered, as they reflect only a subset of the job market dynamics of their time. This paper identifies these biases, quantifies their impact, and proposes solutions for steps from corpus selection to data preparation for subsequent text-mining and analysis. We discuss and exemplify the implications of these biases on research outcomes and suggest methodological adjustments to mitigate their effects, ensuring more reliable insights into the historical labor market. Also, we make a dataset of 15 000 manually annotated ground-truth data available as part of this paper.

Keywords

digitized newspapers, historical job advertisements, historical labour market, data criticism, optical character recognition, page segmentation, post-processing, ground truth

1. Introduction

Digitized newspapers as a source of data bring many opportunities, however, also many challenges and pitfalls. In the JobAds project (FWF P35783), we investigate the evolution of the Austrian labor market through historical job advertisements from digitized newspapers between 1850-1950. Through their analysis, we aim to get insights into the regional and temporal trends in positions offered and sought for, the skills and qualifications required and offered, media strategies, but also social aspects such as gender-specific job offers.

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark *Corresponding author.

[☆] klara.venglarova@uni-graz.at (K. Venglarova); raven.adam@uni-graz.at (R. Adam);

wiltrud.moelzer@uni-graz.at (W. Mölzer); saranya.balasubramanian@uni-graz.at (S. Balasubramanian); joern.kleinert@uni-graz.at (J. Kleinert); manfred.fuellsack@uni-graz.at (M. Füllsack); georg.vogeler@uni-graz.at

⁽G. Vogeler)

^{© 0009-0007-6441-7795 (}K. Venglarova); 0000-0001-7841-2601 (R. Adam); 0009-0002-9517-4531 (W. Mölzer); 0000-0001-7516-7671 (S. Balasubramanian); 0000-0002-1167-9245 (J. Kleinert); 0000-0002-7772-4061 (M. Füllsack); 0000-0002-1726-1712 (G. Vogeler)

^{© 2024} Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Our project aims to cover as much as possible of the Austrian labor market in the defined time period. In this paper, we discuss the difficulties which typically arise when one tries to access data from digitized newspapers. We refer to these difficulties as 'biases', as they have a great potential to skew results and their interpretation.

We used 29 periodics (see Appendix) with the largest number of editions within our time span from the ANNO corpus [16]. From each newspaper, we preselected pages containing job advertisements and converted them to machine-readable text through page segmentation, optical character recognition (OCR), and post-correction.

Each processing step introduces potential biases: selection of periodicals may not be representative, a classifier may systematically misclassify pages containing certain types of ads, segmentation quality may vary with layout complexity or scan quality, OCR may be affected by font use or style particularities such as white letters on a black background, and post-correction effectiveness may vary between newspapers or years because of different typical mistakes.

Biases also arise from the historical context of newspapers, which were only one of several channels realizing matches between job seekers and vacancies [25]. Some jobs were advertised through newspapers more often than others, as in a short radius, it was harder to find highly specialized employees than e.g. unqualified workers who could be placed through a bourse system [13].

This paper addresses how we confront a research question with the technical and historical reality, from corpus selection to OCRed text post-processing. Section 2 describes related work, section 3 provides details about dataset creation. Sections 4 and 5 discuss biases arising from historical context and processing steps respectively, and section 6 outlines future work and concludes the paper.

2. Related Work

Oberbichler and Pfanzelter [17] discuss a large number of 'biases that come along with the processing and datafication of historical newspapers' (p.127) and illustrate them in a case study about return migration. In a keyword-based search, the first challenge is to select the right terms, which is hindered by word flexions and spelling variants, but also by semantic uncertainty, e.g. false positive hits if the term is too broad or not finding all occurrences if the term is too narrow. They show how missing data and varying OCR skew absolute frequencies, and use therefore relative frequencies instead. They also propose improving OCR quality with tools like Transkribus [18] and argue for adding metadata, contextualization of the source documents, providing information about the limitations of the collection, and additional tools in the interfaces.

Wijfjes [28] discusses the relationship between traditional humanities and new digital methods. The most prominent obstacle in using digitized newspapers as a research source is the incompleteness of digitized collections, caused by factors such as costs, time or copyright issues. Relying on the available collections regardless of their broader context can result in working with a 'convenience sample' (p.16), closely related to 'digital laziness' (p.10) which arises from an overreliance on easily accessible digital information. Author also mentions unreliability in OCR and the need for 'complete and uniform data' (p.21). The errors in Optical Layout Recognition (OLR) and Optical Character Recognition (OCR) are crucial problems in machine-readable text creation that have been addressed by several scholars ([23]; [27]; [8]; [6]; [21]). Noisy OCR or its varying quality across the corpus poses problems not only in a keyword search, but also in subsequent tasks, such as Part-of-Speech (POS) tagging, dependency parsing, Named Entity Recognition (NER) or topic modeling. Therefore, it is necessary to assess the influence of the OCR quality on the outcomes of NLP tasks, as discussed e.g. by ([22]; [20]; [26]).

Cordell [4] largely discusses digitized collection in a broader context of the process of digitalization and decisions leading to which material shall be digitized. He distinguishes between printed and digital editions and argues for taking digitized text 'seriously within its own medium' (p.217). On the example of the *Raven* by E. A. Poe, he shows how results of keyword search are dependent on the quality of the OCR output.

We are aware of all these and handle the challenges where they occur. For example, working with relative frequencies of job titles can help to capture the changing demands and offers of a specific position instead of unintentional identification of changing trends of advertising jobs in newspapers. Considering the previous research and our own experience, we search for strategies for mitigating the bias in our data.

3. Dataset

Our dataset comprises 29 newspaper titles from the ANNO corpus, a collection of digitized newspapers provided by the Austrian National Library, from 1850-1950. The 29 newspapers were selected based on the largest number of editions, with a minimal issue period of 10 years. These newspapers are predominantly in German, containing minor numbers of ads in French, English, Czech, Hungarian, Italian and other languages. Because of the time needed for processing the pages, we initially preselected pages containing job advertisements manually, based on observed patterns in the advertisement sections appearance, and later refined the preselection automatically using a transformer-based preselection.

From the preselected pages, we randomly sampled one page per year for each newspaper available for that year, resulting in 3 300 pages. On these pages, we manually annotated all job advertisements using doccano software [15], resulting in 14 985 annotated job ads. These annotations serve us as ground-truth data. The annotated ads were OCRed with the frak2021 model [10] and manually corrected using Transkribus.

4. Biases from Historical Context

In this section, we explore biases arising from the historical context and the nature of newspapers as a medium. As these biases are beyond direct control, we have to gather comprehensive information to adjust our research questions based on available data.

4.1. Newspapers as a Medium

The bias in our research does not start with the selection of the 29 newspaper titles; it starts with the decision to use newspapers as our primary data source to understand labor market development. The matches between the job seekers and vacancies were realized through several channels [5], with different proportions for various types of jobs.

To address this bias, it is crucial to quantify the extent to which job seekers and vacancies were matched through newspaper ads compared to other channels and to understand which jobs were underrepresented or missing in them. If sufficient facts concerning these aspects cannot be found, it is necessary to redefine the research goals. Instead of aiming to describe the entire labor market, we can narrow our scope to explore specific job categories, time periods, or regions.

For our period, approximately 30% of the matches between job seekers and employers were realized through job advertisements in newspapers [13]. Other channels to find matches were personal contacts, asking around, bourse system, municipal placement services and commercial brokers, which came increasingly under administrative control to reduce abuse and were finally replaced by the public job services. Search via newspaper ads was dominating among white-collar jobs [5]. Blue-collar workers were underrepresented in this channel [13]. Based on these facts, we can aim to compare e.g. qualifications offered with requirements presented in job searches and job offers for white-collar workers.

4.2. Newspapers and their focus

The focus of newspapers shall also be considered. To gain a comprehensive understanding of our data, we should address the following questions regarding the newspapers:

- What is the political orientation of the newspapers?
- Which geographical area do they cover?
- What is their temporal coverage?
- What is their social focus?
- Who are the intended readers?

Access to such meta-information about newspapers provides crucial context and facts. While no single newspaper can fully represent the labor market, a selection of heterogeneous titles covering various aspects can collectively offer a broader perspective. Our selected newspapers were issued in several geographical regions (e.g. *Arbeiterwille* in Graz, *Arbeiter Zeitung* in Wien, *Linzer Volksblatt* in Linz, *Salzburger Chronik für Stadt und Land* in Salzburg), covering longer periods (see Fig. 1), most of them daily newspaper but also with focus on travels (*Fremden-Blatt*) or workers (*Arbeiterwille*). We included newspapers with various orientations, such as social-democratic (e.g. *Arbeiterwille*, *Arbeiterzeitung*), liberal-democratic (*Prager Tagblatt*), or nationalistic (*Salzburger Volksblatt: die unabhängige Tageszeitung für Stadt und Land Salzburg*). However, achieving an ideal representation still remains extremely challenging due to realworld complexities.



Figure 1: Temporal distribution of selected newspaper titles according to their issuance years. Data source: [16].

4.3. Who Advertises in Newspapers?

To draw unbiased conclusions from job advertisements about the entire labor market, the underlying assumption would be that people offering and searching through job ads (or the job ads that we analyze) have the same characteristics of interest as job offerers or seekers who do not use newspaper ads. Since we cannot often know that, we need to ask who actually advertises in newspapers.

For instance, in a small town with only one factory, there may be no need to advertise for unqualified workers in newspapers. Similarly, workers in such towns may not need to advertise either, as they can rely on local announcements or neighbors or family. Consequently, these types of ads may be missing or underrepresented in our corpus. On the contrary, if the factory seeks highly specialized personnel, they may expand their outreach by utilizing the newspapers.

A similar scenario applies to a city baker. If they seek someone locally, they might rely on the people in their surrounding. However, if they require an apprentice from the countryside, such as e.g. in (Fig. 2), they would more likely advertise in newspapers.



Figure 2: Job advertisement for a baker apprentice from the countryside. Source: [2].

This highlights that in our corpus, advertisements might be overrepresented that involve a greater distance, e.g. geographical distance (a baker from a city seeks an apprentice from the countryside) or social distance (higher class person looking for a servant). In other cases, despite a short geographical distance in large cities, complexity of an urban society and anonymity plays a role. Also, the ads in our corpus may be extreme from some point of view, e.g. seeking a highly qualified person, a factory with bad working conditions, high demand for workers for seasonal work, a person who struggles to find a job for a longer time. On the other hand, some ads may be missing, e.g. from people who could not afford to pay for an ad in newspapers.

5. Biases from Data Processing

This section describes biases that data processing introduces, and presents strategies to mitigate them. Our processing pipeline contains steps from corpus selection to the cleaned data, which can be used for meaningful economical analysis.

5.1. Corpus creation

The initial step involves building a corpus. While in the ideal case, researchers would work with the entire population of data, practical limitations in terms of time and computation power or missing newspapers in digitized collections make this only hardly feasible.

We come as close as we can to the entire population given our resource constraints by selecting the newspapers with the most editions that were issued for at least 10 years, which allows for a comparison in time. In a corpus selection, attention needs to be paid to include heterogeneous newspapers that will cover a wide scope. For further information, see subsection 4.1.

5.2. Preselection of Relevant Pages

Initially, we manually preselected pages likely to contain job advertisements by examining sampled issues from all newspapers and identifying patterns in the job ad sections. In a later stage, we refined this process using a transformer-based model for the same task, by fine-tuning the microsoft/dit-large-finetuned-rvlcdip model [9] on our data. The fine-tuned model reached a f1 score of 0.88 and recall of 0.89 on the testing data.

Preselection can be a dangerous process that can lead to excluding relevant data and bias the results. Our strategy against this pitfall was aiming for the highest recall possible, favoring the inclusion of non-relevant pages over the exclusion of relevant ones. Our model's results, which indicated that only about 34% of the preselected 4,000,000 pages actually contained job advertisements, give us confidence that we have captured most relevant pages.

5.3. Segmentation process

Page segmentation is a step which has a direct impact on the OCR quality ([11]; [12]; [1]; [3]). However, segmentation quality is often assessed through subjective visual control [7], which does not offer complex insights into segmented data quality.

To address this task, we adopt a methodology for segmentation evaluation from [24]. This method is not based only on the area of intersection between the annotated and predicted region, as this can be skewed if graphical elements or large blank spaces are present in data. Instead, it also benefits from information about the presence of the text in the non-intersecting parts of the predicted region and its ground-truth.

We manually annotated nearly 15 000 job ads from our corpus to create ground truth data. We make the annotated data publicly available in the GitHub repository:

https://github.com/JobAds-FWFProject/Ground-Truth-CHR2024. This allows us to identify newspapers with lower segmentation accuracy and generate additional ground-truth data to fine-tune our segmentation algorithm effectively. The results of the ongoing work on the evaluation of the segmentation quality will be published in another publication.

5.4. OCR Quality

OCR quality is the most prominent source of bias, as highlighted in prior research ([17]; [23]; [27]). Variations in OCR accuracy can lead to discrepancies in e.g. keyword searches and affect data reliability.

The first step to mitigate this bias is quantifying the OCR quality by e.g. a character error rate (see Fig. 3). Although an approximation of the OCR quality can be obtained by checking words against a dictionary [21], we decided to manually check and correct a sample of the advertisements. This cleaned data serves us to (1) quantify the quality of the OCR, (2) provide us with high-quality data for text-mining experiments and (3) give us information about the most common mistakes in recognition which we can use for automatic post-corrections. A pure dictionary-based approach would introduce a bias by words missing in the dictionary, e.g. abbreviations or names, which both are often in our datasets. Based on our manually transcribed ground-truth, our OCR reaches a SacreBLEU score of 67.5%, word error rate of

30.6% and character error rate of 12.2%. However, apart from this overall evaluation, it is crucial to compare the OCR quality across newspapers and years.



Figure 3: Character Error Rate of the OCRed text in Prager Abendblatt.

5.5. Post-correction

Post-correction is the last step in our pipeline that affects data quality. Starting with a varying OCR quality across newspaper titles or years, the post-correction can both reduce or amplify the discrepancies in data quality. To quantify this problem, we measure the error rate after different post-correction steps on the samples that were manually checked and corrected. As this sample contains ads across all newspaper titles and years, we will be able to compare text quality before and after the post-correction step.

For the post-correction process, we fine-tune the hmbyt5-preliminary model¹ on the IDCAR2019-POCR dataset for OCR correction [19], which significantly improves text accuracy. On the IDCAR data, we reach the SacreBLEU score of 72.25% compared to only 10.83% achieved by the original OCRed text.

While the model significantly improves the quality of a very bad OCR, our OCR already reaches a higher score (67.5%) before the post-correction step. As there is a large discrepancy between the OCR quality of the training data and our OCR data, the post-correction model itself sometimes introduces new errors (see Tab 1). Some mistakes concern only interpunction, some change letters leading to words with no meaning, but some of them introduce a semantically different result, such as painters (Malerinnen) from ironers (Büglerinnen).

¹hmbyt5-preliminary model [18.6.2024]

Ground Truth	OCR Post-Correction		
Geübte Büglerinnen und	Geübte Bualerinnen und	Geübte Malerinnen und	
Lehrmädchen	Lehrmädchen	rmädchen Lehrmädchen	
Büglerinnen und Lehrmäd- Büglerinnen und Lehrmäd- Züglerinnen und Lehrmäd-			
chen auf neue Herrenhemden	chen auf neue Herrenhemden	chen auf neue Herrenhemden	

 Table 1

 Mistakes introduced through the post-processing step on two text examples.

We also need to consider that not all characters can be corrected through post-processing. For example, for incorrectly recognized numbers, the original information is lost and cannot be gained back by post-processing (see Fig. 4). This potentially affects our ability to analyze details about e.g. offered salary where available.

Seite 38. Nr. 148. JoosbruSer Nachrichten Samstag den I^Luli 1911. Gebe einem sehr geehrten Publikum bekannt, daß ich ab K. Juli mein Seite 38. Nr. 148. Innsbrucker Nachrichten Samstag den 1. Juli 1911. Gebe einem sehr geehrten Publikum bekannt, daß ich ab 6. Juli mein

Figure 4: Example of the original and post-corrected text. Although the 'K' was turned into a number, the correct character is the '1' and not the '6'. Source: [14].

5.6. Example: Looking for a Paperhanger

To illustrate how technical biases can affect historical data interpretation, we examine the demand for five different positions by comparing the frequency of job ads mentioning them between 1850-1900 and 1901-1950. We use a sample of 2779 job ads with both raw OCR and manually corrected transcriptions. These ads are divided into two data sets: one for 1850-1900 and another for 1901-1950.

First, we compare the absolute frequencies of the positions in both OCRed text, with the results in Tab. 2. E.g., the 'Tapezierer' (paperhanger) appears 2, resp. 4 times in the data sets.

Table 2

Absolute frequencies of positions found in the OCRed text between 1850-1900 and 1901-1950.

	1850-1900	1901-1950
Tapezierer (Paperhanger)	2	4
Stubenmädchen (Maid)	5	25
Verkäuferin (Shop Assistant f.)	7	14
Bäcker (Baker)	4	22
Vertreter (Agent/Representative)	6	18

For the example of the paperhanger, the OCR data suggests a low frequency of ads, with a slight increase in the 20th century. This might lead to the conclusion that paperhanger jobs were rare and that their demand doubled in the first half of the 20th century. However, when we look at the manually corrected versions of the text, we obtain the following results (Tab. 3):

Manually corrected data also indicates an increase in demand over time. However, absolute frequencies can more reflect the amount of available data rather then trends of employment.

Table 3

	1850-1900	1901-1950
Tapezierer (Paperhanger)	6	10
Stubenmädchen (Maid)	7	28
Verkäuferin (Shop Assistant f.)	8	21
Bäcker (Baker)	4	37
Vertreter (Agent/Representative)	10	21

Absolute frequencies of positions found in the manually corrected text between 1850-1900 and 1901-1950.

To account for this, we divide absolute frequencies by the number of ads available for these time periods in our sample, and we get another information (Tab. 4):

Table 4

Relative frequencies of positions found in the manually corrected text between 1850-1900 and 1901-1950.

	1850-1900	1901-1950
Number of job ads	1016	1763
Tapezierer (Paperhanger)	0.591%	0.567%
Stubenmädchen (Maid)	0.689%	1.588%
Verkäuferin (Shop Assistant f.)	0.787%	1.191%
Bäcker (Baker)	0.394%	2.099%
Vertreter (Agent/Representative)	0.984%	1.191%

When adjusted for the total number of ads, the relative frequency of paperhanger ads shows that the demand remained almost constant. The apparent increase in absolute numbers is due to the larger volume of data available for the later period, not an actual rise in demand. The example of the paperhanger, and similarly of the other positions, demonstrate the need for cautious interpretation of historical job ads, and that the OCR errors do not impact all positions in the same way. Although similar tests need to be done on a larger scale, this example illustrates how easily we can run into a false interpretation.

6. Conclusion

In this paper, we addressed the biases encountered in our research of the labor market based on job advertisements from digitized newspapers. Our investigation focused on two sources of bias: those arising from the historical context of the newspapers and those introduced during data processing. The first set of biases stems from the nature of newspapers as a data source. This includes the selection of newspaper titles, their political and social orientations, geographic reach, and the profiles of advertisers and readers. These factors shape the type of job ads available, leading to over- or underrepresentation of certain job types and demographic groups. We emphasize that understanding these contextual elements is crucial. Researchers must adjust their research questions to align with the data's actual scope and representation rather than assuming a comprehensive view of the labor market.

The second set of biases arises from the various stages of data processing. This encompasses corpus creation, pre-selection of relevant pages, page segmentation, OCR quality, and post-correction processes. Each stage presents potential pitfalls that can skew results. We highlighted the importance of high recall in data preselection, evaluating segmentation and OCR accuracy, and manual corrections and error analysis. Through a practical example, we demonstrated how technical biases can skew historical job market analyses.

Our ongoing work includes gathering meta-information about each newspaper title and the labor market to better understand the matching of job applicants with vacancies. Also, we continue with further manual corrections of OCRed text to ensure consistent data quality across our corpus. We plan to deeper investigate the post-processing aspects, such as the model's skills transferability in cases, when there is a great difference between the OCR quality of the training and testing data. Additionally, we plan to implement strategies like data augmentation for underrepresented job categories.

Data Availability

The annotated ground truth data is publicly available at: https://github.com/JobAds-FWFProject/Ground-Truth-CHR2024.

Acknowledgments

We thank the Austrian National Library (ÖNB) for providing the data. We would also like to thank Meike Linnewedel, Clara Hochreiter and Melanie Frauendorfer for their efforts in correcting and annotating the ground truth. This work was supported by the FWF under grant number P 35783.

A. Appendix. List of newspaper titles present in our corpus

Arbeiterwille Arbeiter-Zeitung Bregenzer Tagblatt/Vorarlberger Tagblatt Das Vaterland Deutsches Volksblatt Die Presse Freie Stimmen Fremden-Blatt Grazer Tagblatt Grazer Volksblatt Illustrierte Kronen Zeitung Innsbrucker Nachrichten Linzer Tages-Post Linzer Volksblatt Morgen-Post Neue Freie Presse Neues Wiener Journal: unparteiisches Tagblatt Neues Wiener Tagblatt (Tagesausgabe) Neuigkeits - Welt - Blatt Pester Lloyd Pilsner Tagblatt Prager Abendblatt Prager Tagblatt Reichspost Salzburger Chronik fuer Stadt und Land Salzburger Volksblatt: die unabhaengige Tageszeitung fuer Stadt und Land Salzburg Vorarlberger Landes-Zeitung Vorarlberger Volks-Blatt Wiener Zeitung

References

- R. Barman, M. Ehrmann, S. Clematide, S. Oliveira, and F. Kaplan. "Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers". In: *Journal of Data Mining & Digital Humanities* HistoInformatics (2021). DOI: 10.46298/jdmdh.6107.
- [2] (. W. Blatt. 30.7.1921. 1921. URL: https://anno.onb.ac.at/cgi-content/anno?aid=nwb%5 C&datum=19210730%5C&seite=8.
- [3] Y. Can and M. Kabadayi. "CNN-Based Page Segmentation and Object Classification for Counting Population in Ottoman Archival Documentation". In: *Journal of Imaging* 6 (2020), p. 32. DOI: 10.3390/jimaging6050032.
- [4] R. Cordell. "" Q i-jtb the Raven": Taking Dirty OCR Seriously". In: Book History 20.1 (2017), pp. 188–225.
- [5] A. Faust. "Arbeitsmarktpolitik im deutschen Kaiserreich: Arbeitsvermittlung, Arbeitsbeschaffung und Arbeitslosenunterstützung 1890-1918". In: (*No Title*) (1986).
- [6] J. Jarlbrink and P. Snickars. "Cultural heritage as digital noise: nineteenth century newspapers in the digital archive". In: *Journal of Documentation* 73.6 (2017), pp. 1228–1243. DOI: 10.1108/jd-09-2016-0106. URL: https://doi.org/10.1108/JD-09-2016-0106.
- [7] X. Jiang, C. Marti, C. Irniger, and H. Bunke. "Distance Measures for Image Segmentation Evaluation". In: *EURASIP Journal on Advances in Signal Processing* 2006.1 (2006), p. 035909. DOI: 10.1155/asp/2006/35909. URL: https://doi.org/10.1155/ASP/2006/35909.
- [8] E. Late and S. Kumpulainen. "Interacting with digitised historical newspapers: understanding the use of digital surrogates as primary sources". In: *Journal of Documentation* ahead-of-print (2021). DOI: 10.1108/jd-04-2021-0078.

- [9] D. D. Lewis, G. Agam, S. E. Argamon, O. Frieder, D. A. Grossman, and J. Heard. "Building a test collection for complex document information processing". In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (2006).
- [10] M. U. Library. *frak2021*. Version frak2021-0.905. 2021. URL: https://ub-backup.bib.uni-ma nnheim.de/~stweil/tesstrain/frak2021/tessdata%5C%5Fbest/frak2021-0.905.traineddata.
- [11] B. Liebl and M. Burghardt. "An Evaluation of DNN Architectures for Page Segmentation of Historical Newspapers". In: 2020 25th International Conference on Pattern Recognition (ICPR). 2021, pp. 5153–5160. DOI: 10.1109/icpr48806.2021.9412571.
- [12] J. Martínek, L. Lenc, and P. Král. "Building an efficient OCR system for historical documents with little training data". In: *Neural Computing and Applications* 32.23 (2020), pp. 17209–17227. DOI: 10.1007/s00521-020-04910-x. URL: https://doi.org/10.1007/s00521-020-04910-x.
- [13] W. Mölzer and J. Kleinert. Emergence of the Austrian labor market. 2024. URL: https://sta tic.uni-graz.at/fileadmin/%5C%5Ffiles/%5C%5Fproject%5C%5Fsites/%5C%5Fhistorical-jo b-ads/Emergence%5C%5FAustrian%5C%5Flabor%5C%5Fmarket.pdf.
- [14] I. Nachrichten. 1.7.1911. 1911. URL: https://anno.onb.ac.at/cgi-%20content/anno?aid=ibn %5C&datum=19110701%5C&seite=38.
- [15] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, and X. Liang. doccano: Text Annotation Tool for Human. 2018. URL: https://github.com/doccano/doccano.
- [16] Ö. Nationalbibliothek. ANNO Historische Zeitungen und Zeitschriften. 2021. URL: https://a nno.onb.ac.at/.
- [17] S. Oberbichler and E. Pfanzelter. "Tracing Discourses in Digital Newspaper Collections". In: *Digitised Newspapers – A New Eldorado for Historians?* Ed. by E. Bunout, M. Ehrmann, and F. Clavert. De Gruyter Oldenbourg, 2023, pp. 125–152. DOI: 10.1515/9783110729214-007. URL: https://doi.org/10.1515/9783110729214-007.
- [18] P. Kahle, S. Colutto, G. Hackl, and G. Mühlberger. "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). Vol. 04. 2017, pp. 19–24. DOI: 10.1109/icdar.2017.307.
- [19] C. Rigaud, A. Doucet, M. Coustaty, and J.-P. Moreux. "ICDAR 2019 Competition on Post-OCR Text Correction". In: Proceedings of the 15th International Conference on Document Analysis and Recognition (2019). 2019, pp. 1588–1593.
- [20] K. J. Rodriguez, M. Bryant, T. Blanke, and M. Luszczynska. Comparison of Named Entity Recognition tools for raw OCR text. 2012. DOI: 10.13140/2.1.2850.3045.
- [21] C. Strange, D. McNamara, J. Wodak, and I. Wood. "Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers". In: *digital humanities quarterly* 8 (2014).

- [22] D. Strien, K. Beelen, M. Coll Ardanuy, K. Hosseini, B. Mcgillivray, and G. Colavizza. Assessing the Impact of OCR Quality on Downstream NLP Tasks. SciTePress, 2020. DOI: 10.5 220/0009169004840496.
- [23] A. Torget. "Mapping Texts: Examining the Effects of OCR Noise on Historical Newspaper Collections". In: *Digitised Newspapers – A New Eldorado for Historians?* Ed. by E. Bunout, M. Ehrmann, and F. Clavert. De Gruyter Oldenbourg, 2023, pp. 47–66. DOI: 10.1515/978 3110729214-003.
- [24] K. Venglarova, R. Adam, S. Balasubramanian, and G. Vogeler. "Quantifying Page Segmentation Quality in Historical Job Advertisements Retrieval". 2024. URL: https://inria.h al.science/hal-04560463.
- [25] S. Wadauer, T. Buchner, and A. Mejstrik. "The Making of Public Labour Intermediation: Job Search, Job Placement, and the State in Europe, 1880–1940". In: *International Review* of Social History 57 (S20 2012), pp. 161–189. DOI: 10.1017/s002085901200048x.
- [26] D. Walker, W. Lund, and E. Ringger. *Evaluating Models of Latent Document Semantics in the Presence of OCR Errors*. Association for Computational Linguistics, 2010. 240 pp.
- [27] M. Wevers. "Mining Historical Advertisements in Digitised Newspapers". In: Digitised Newspapers – A New Eldorado for Historians? Ed. by E. Bunout, M. Ehrmann, and F. Clavert. De Gruyter Oldenbourg, 2023, pp. 227–252. DOI: 10.1515/9783110729214-011.
- [28] H. Wijfjes. "Digital Humanities and Media History: A Challenge for Historical Newspaper Research1". In: *TMG Journal for Media History* (2017). DOI: 10.18146/2213-7653.2017 .277.