

Processing Digitized Text on an Example of Job Advertisements from Austrian Periodicals from 1850-1950

Klara Venglarova¹
¹University of Graz
klara.venglarova@uni-graz.at



Corpus of Digitized Newspapers

- ANNO Corpus (Austrian National Library, 2021)
- 1850-1950
- 29 Newspaper Titles
- Predominantly German-language job ads
- 3355 pages, manually annotated 15 000 advertisements (Team effort)



Fig. 1: Job advertisement for a baker apprentice from the countryside. Source: (Neuigkeits) Welt Blatt, 30.7.1921, p. 8. <https://anno.onb.ac.at/cgi-content/anno?aid=nwb&datum=19210730&seite=8> [20.6.2024]

From Image to Text

- Image pre-processing (Otsu, 1979)
- Page segmentation – eynollah (Rezanezhad et al., 2023)
- OCR – frak2021 (Mannheim University Library, 2021): mean CER 15.5%, median CER 7.3%

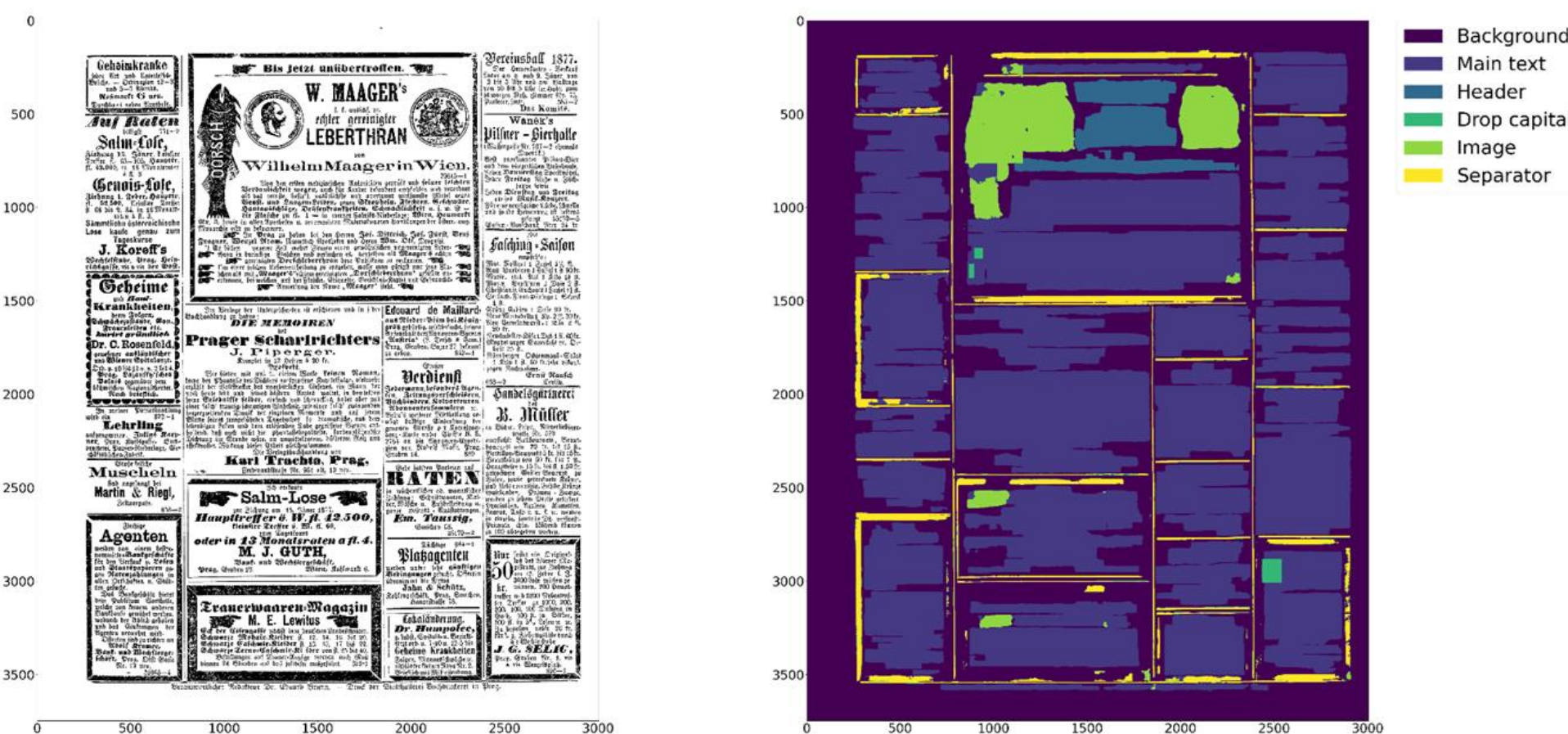


Fig. 2: Layout detection on a newspaper page with Eynollah software. Source: Prager Abendblatt, 10.1.1877, p.4, <https://anno.onb.ac.at/cgi-content/anno?aid=pab&datum=18770110&seite=4> [cited 2.5.2025].

Text Type Classification

- Job Ad vs. Non-Job Ad Classification
- <https://huggingface.co/stefan-it/zeitungs-lm-v1> [cited 13.5.2025]

Method	AUC-PR
zeitungs-lm-v1 (fine-tuning for classification task)	0.9692
bert-base-historical-german-rw-cased (fine-tuning for classification task)	0.9532
bert-base-german-cased (fine-tuning for classification task)	0.9476
jobBERT-de (fine-tuning for classification task)	0.9392
TF-IDF (SVM)	0.8953
Unigrams + Bigrams (SVM)	0.8692
Unigrams + Bigrams + Trigrams (SVM)	0.8672
Bag-of-Words from Verbs and Nouns - without Lemmatization (SVM)	0.8182

Tab. 1: AUC-PR for job advertisements identification on the testing dataset.

Job Titles Extraction

- The first step for an economical analysis
- Dictionary-based method – HISCO (Leeuwen et al., 2002)
- Rule-based – POS and DEP, sentence structure
- NER – fine-tuning a transformer model using SpaCy’s (Honnibal and Montani, 2017) framework
- Text Generation – a translation task

Eine **Naturblumenbinderin** **position** gelernt in Wien, bittet um Kondition.
Emma Ziegler, Gärtnergasse 7. Perchtoldsdarf. N.ºOc. 3123—6

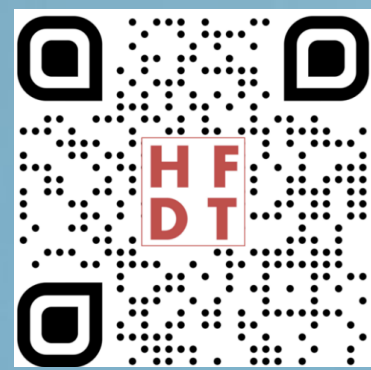
Method	Raw OCR		Automatically Post-corrected Text		Human-corrected Text	
	Train	Test	Train	Test	Train	Test
Dictionary-based	0.517	0.509	0.554	0.552	0.589	0.582
Dictionary-based with Tolerance	0.385	0.379	0.395	0.397	-	-
Rule-based	0.668	0.65	0.681	0.722	0.743	0.714
NER (CPU)	0.866	0.776	0.932	0.842	0.995	0.88
NER (GPU)	0.882	0.855	0.942	0.907	0.996	0.956
Text Generation	0.802	0.754	0.829	0.775	0.990	0.922

Tab. 2: F1-scores for position extraction methods on the texts of different quality.



www.clarin.eu

clarin@clarin.eu
 github.com/clarin-eric



Österreichische Nationalbibliothek



Österreichischer Wissenschaftsfonds



CLARIN ERIC was established in 2012 and received ESFRI Landmark status in 2016