

Behind the Labels: Manually Annotating 100 Years of Historical Job Advertisements to Develop Automatic Semantic Annotation Models

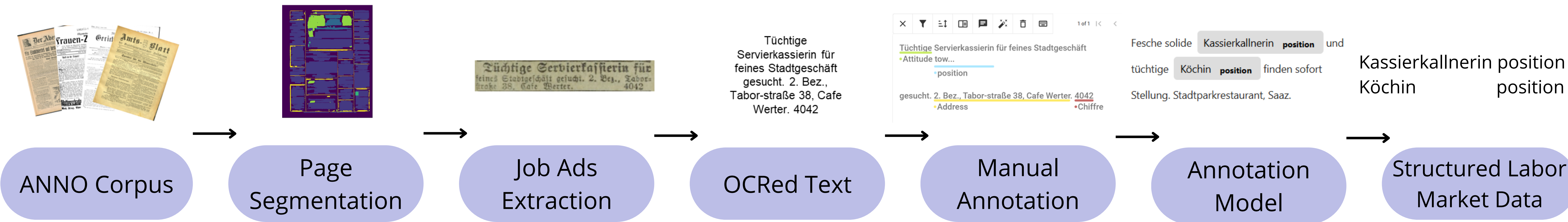
Klara Venglarova, Raven Adam, Wiltrud Mölzer, Jörn Kleinert, Georg Vogeler
University of Graz

MOTIVATION

- Historical job advertisements are valuable for studying:
 - Labour markets over time
 - Occupational change
 - Skills and qualifications
 - Salaries
 - Gendered labour markets



<https://anno.onb.ac.at/cgi-content/anno?aid=nwb&datum=19210730&seite=8> [7.7.2026]



ANNOTATION DATASET

- 14,985 manually annotated job advertisements
- Austrian newspapers
- 1850–1950
- Representative stratified sample
- Annotated in Doccano (Nakayama et al., 2018)

WHAT WAS ANNOTATED:

Layout Level

- Job offers
- Job searches
- Service offers
- Mediations
- Headings

Chunk Level - Examples:

- Job titles
- Salary
- In-kind salary
- Skills
- Requirements

SOLVING ANNOTATION CHALLENGES

- Iterative guideline development
- Regular annotator discussions
- Continuous refinement
- Context-sensitive decisions
- Consistency checks

RESULTS

- High-quality annotated corpus, annotation guidelines, training dataset for automatic annotation model

Model performance:

Dataset	F1
Human-corrected text	0.95
Unrelated OCR dataset	0.83



Doccano Interface, source: <https://anno.onb.ac.at/cgi-content/anno?aid=gtb&datum=18960101&seite=18> [22.4.2025].

ANNOTATION CHALLENGES

1) Ambiguous syntax (Inversion)

"Perfekte Kontoristin sucht Schuhfabrik"
"Perfect office clerk seeks shoe factory" or "Shoe factory seeks a perfect office clerk" → **Requires contextual reasoning**

2) Missing historical context

"Madelaine Arztesgattin" [doctor's wife] → Person? Occupation? Product? → **Historical knowledge needed**

3) Unclear entity boundaries

"Junger Mann" [young man], "Französin" [French women], "Mädchen" [maid] → not really occupations, **require adding new label** 'not-positions'

High-quality manual annotation is essential for **robust historical semantic annotation**. Careful **guideline development** and **human interpretation** enable reliable automated extraction from noisy historical newspapers.

