AI-supported indexing of handwritten dialect lexis: The pilot study "DWA Austria" as a case study.

Kunzmann, Markus
markus.kunzmann@oeaw.ac.at
Austrian Academy of Sciences (ÖAW), Austria

Traditionally, two approaches have developed in dialectology that focus on researching lexical variation: on the one hand, dialect dictionaries, whose task is to document dialect vocabulary; on the other hand, dialect atlases, whose focus is on linguistic-geographical variation in dialect vocabulary. The *Wörterbuch der bairischen Mundarten in Österreich* (WBÖ), a long-term project of the Austrian Academy of Sciences (ÖAW), is an undertaking of the first type. Until 2015, the first five volumes (*A–Ezzes*) were published as printed works; since 2018, the *Lexikalisches Informationssystem Österreich*/*Lexical Information System Austria* (LIÖ) has served as the publication platform for the articles starting with the letter *F*. LIÖ is a cooperation project between the *Austrian Centre for Digital Humanities and Cultural Heritage* (ACDH-CH) of the ÖAW and the FWF Special Research Programme *German in Austria* (DiÖ) of the University of Vienna. Now, the information system only contains content related to the WBÖ project itself. As a lexically oriented platform, however, its content is to be expanded in the coming years, i.e., lexical material from other corpora is also to be made accessible via it.

A first expansion of LIÖ will take place in October 2022 with the pilot study *DWA Austria*, a cooperation project between the *Research Center Deutscher Sprachatlas* (DSA) of the Philipps University of Marburg and the Department of Linguistics of the ACDH-CH. Within the framework of this cooperation, the entire Austrian surveys of the *Deutscher Wortatlas*/*German Word Atlas* (DWA) are to be digitally processed for the first time. Finally, the project will serve to expand the paradigm for researching lexical variation described above to include a dialect-geographical component. In this collaboration, the Marburg team will provide the high-resolution scans of the DWA surveys. The team in Vienna is building a model for automatic transliteration on this basis with the help of the *Transkribus* software, which in turn can be used by the DSA team for the German DWA sheets.

The surveys for the DWA were conducted indirectly between 1939 and 1942 and are still among the most comprehensive surveys of the 20th century. Questionnaires were sent to a total of about 50,000 places, 3,700 of which are in the territory of the Republic of Austria. Since previous automatic text recognition methods (OCR) have only provided insufficient results, the questionnaires, which were mostly handwritten, had to be laboriously transliterated manually. The recent use of artificial intelligence (AI) has been showing promising results for several years.

With the help of the *Transkribus* platform, the Austrian DWA questionnaires are now being captured and made usable as part of a pilot study. Unlike conventional OCR products, *Transkribus* uses artificial intelligence (AI) to convert the written content of digital records into searchable text. The scans of the DWA sheets were made by the DSA and made available to the ACDH-CH. There, in a first step, they manually transliterate a set of scanned sheets. This step can be supported by already existing models that are, for example, tailored to German Kurrent script. Based on these correct transliterations and the corresponding scans, a model can now be built with the help of Deep Learning that is tailored to the document type. The layout of the text is also considered. First models on the minimum amount of training material still showed a rather high error rate (CER Val. 9.4) and have been continuously improved since then.

The pilot study *DWA Austria* shows how data sets that could previously only be used to a limited extent due to time-consuming and costly efforts can now be opened by AI-supported methods to an extent that would not have been possible with conventional methods. In particular, the example shows how the respective expertise of the individual project partners can bring about a significant increase in efficiency and thus once again illustrates the potential for synergies that result from cooperation.

Bibliography

DiÖ = SFB German in Austria. URL: https://www.dioe.at/en/ [2023-05-01]

DSA = Forschungszentrum Deutscher Sprachatlas. URL: https://www.uni-marburg.de/en/fb09/dsa [2023-05-01]

DWA = Mitzka, Walther & Ludwig Erich Schmitt. 1951 – 1980. Deutscher Wortatlas. Gießen: Schmitz

LiÖ = Lexikalisches Informationssystem Österreich. URL: https://lioe.dioe.at/ [2023-05-01]

Transkribus. AI powered Handwritten Text Recognition. URL: https://readcoop.eu/transkribus/ [2023-05-01]

WBÖ = Österreichischen Akademie der Wissenschaften.1970 – 2015. Wörterbuch der bairischen Mundarten in Österreich. Wien: Verlag der Österreichischen Akademie der Wissenschaften.