From paper slips via TUSTEP to TEI – creating a lexicographic information system from the WBÖ database

The Dictionary of Bavarian Dialects in Austria (Wörterbuch der bairischen Mundarten in Österreich 'WBÖ') is a long-term project, whose main goal is the comprehensive lexicographic documentation of the manifold Bavarian base dialects in Austria and South Tyrol. The data used for writing the dictionary articles where recorded by so-called "Sammler" (gatherers) on paper-slips and collected in the "Hauptkatalog" (main catalogue). Since 2016, it is (re)located at the department "Variation and Change of German in Austria" at the Austrian Centre for Digital Humanities (ACDH) at the Austrian Academy for Sciences (ÖAW).

With the main purpose of facilitating and accelerating the process of writing dictionary articles, the hand-written paper slips were entered manually into a TUSTEP system and, subsequently, converted into TEI-XML. Altogether, the database contains about 2.4 million entries. Additionally, the original paper slips are being scanned and will be linked to the corresponding data.

In our talk, we want to focus on two closely interrelated topics. The first one concerns the technical problems and challenges resulting from the conversion of the original extremely diverse collection of data into TEI format. This work has offered us some insight as to how well that markup vocabulary is able to handle such a dataset (cf. Bowers 2017).

Our second topic focuses on the conceptual aspects of the future WBÖ work, which will be twofold: on the one hand, we will continue to write "classical" dictionary articles, which will appear in a revised and modernized format. On the other hand, the goal is to create a comprehensive online lexicographic information system, i.e. a (re)search tool which will allow users to perform queries regarding different aspects of the database such as lemma, sense, or geographic location/region. Additionally where existing, supplementary materials such as scans of the original paper slips or scans of questionnaires will also be accessible.

In this presentation, we will: trace the different steps that have already been taken to create our current digital database from the original data collection; discuss the challenges we are still facing with respect to the internal structure and homogeneity of the data; and describe the approaches we are considering to address such challenges.

References:

Bowers, Jack T. (2017): TEI conversion of Bavarian dialect lexical resources: insights, observations, challenges, next steps. Presented at the COST-ENeL 2017, Budapest, Hungary.