A mountain of work. Building an Alpine Heritage Text Corpus

Claudia Posch, Gerhard Rampl and Bettina Larl (University of Innsbruck, Austria)

The proposed poster demonstrates the process of building a large heritage corpus in the German language, which is an ongoing project at the University of Innsbruck. Inspired by the Swiss project Text+Berg digital (Bubenhofer et al., 2015) the Austrian project Alpenwort's is currently working on a POS & NER annotated corpus of alpine heritage texts. The Alpenwort corpus contains 126 yearbooks of the Austrian Alpine Club Magazine (ZAV =Zeitschrift des Deutschen und Österreichischen Alpenvereins) starting from as early as 1869 until 1998. All of the volumes were already digitized and OCRed and are now part of a TEI-conform XML corpus with approximately 18,6 million words. One particular problem the project had to face during the digitization process was the large amount of text in Gothic script (The ZAV was issued in Gothic script from 1914 to 1962). To solve the resulting OCR-errors a range of semiautomated correction steps were developed.

The ZAV is an extraordinarily interesting source because of its continuity but also because of its thematic diversity. In its first decades the magazine contributions reflect the ongoing touristic and cartographic exploration of the Alps and the economic and scientific discoveries involved. During the 20th century perspectives expanded to the mountains of the world. Globally relevant topics such as environment and nature protection are discussed as well as questions of regional identity and cultural heritage.

The proposed poster presentation gives an overlook of the important steps in our efforts in building the Alpenwort Corpus: From scanning more than 42.000 book pages to Optical Character Recognition to logical structure extraction, the correction of structural elements and OCR-correction. We will also discuss the automated data annotation and enrichment, that contains tokenization, POS-tagging as well as named entity recognition.

The Alpenwort corpus will be freely available for the research community later in 2017 as an XMLversion as well as integrated in the tool Hyperbase, which was developed by our project partners in Nice, France.

References

Bubenhofer, Noah & Juliane Schröter (2012). Die Alpen. Sprachgebrauchsgeschichte - Korpuslinguistik - Kulturanalyse . In P. Maitz (ed.), *Historische Sprachwissenschaft: Erkenntnisinteressen, Grundlagenprobleme, Desiderate; (Studia linguistica Germanica 110)* (pp. 263–287). Berlin: de Gruyter.

Bubenhofer, N., Volk, M., Leuenberger, F. & Wüest, D. (2015). *Text+Berg-Korpus (Release 151_v01)*. Institut für Computerlinguistik, Universität Zürich.