A Mountain of Work

Building an Alpine Heritage Text Corpus

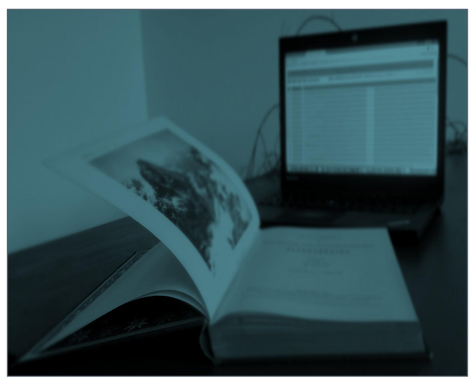
Claudia Posch[1]. Gerhard Rampl[2]. Bettina Larl[3]. Gerald Hiebel[4]. Eva Zangerle[5].

[1-3] Linguistics @ University of Innsbruck (UIBK). [4] Unit for Surveying and Geoinformation @ UIBK. [5] Databases and Information Systems @ UIBK. *Funding: ÖAW go!digital. Partners: Österreichischer Alpenverein, Innsbruck. Abteilung für Digitalisierung & Elektronische Archivierung @ UIBK. Martin Volk, Phillip Ströbel, Noah Bubenhofer, Institut für Computerlinguistik, University of Zürich. Laurent Vanni, Hyperbase Web Edition @ Université de Nice Sophia-Antipolis. Margaret McMahon @ New Zealand Alpine Club, Christchurch.*

INTRODUCTION - THE AUSTRIAN ALPINE CLUB JOURNAL

Since the 1869 the Austrian Alpine Club has been publishing its journal *Zeitschrift des Deutschen und Österreichischen Alpenvereins* (ZAV) in the form of an almanac. The project *Alpenwort* at the University of Innsbruck digitised the ZAV and turned it into a fully POS-tagged thematic corpus, which is integrated in *IMS Open CWB (CQPweb)* as well as in the online platform *Hyperbase*. It will be openly available for the research community via these platforms and in XML format by October 2017.

BUILDING THE CORPUS

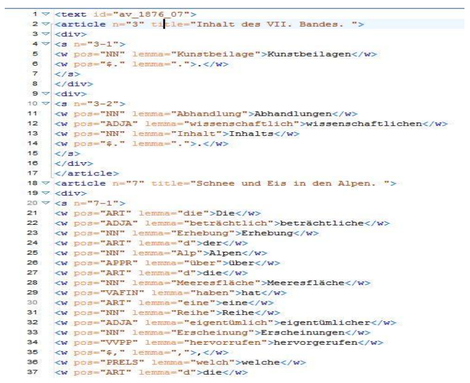


SCANNING AND OCR

The project includes all ZAV-volumes from 1869 – 1998. On average they have 300 pages (min. 108 - max 848), altogether 42118 pages were scanned and OCRed with ABBYY fine reader. The books cover a rather diverse range of articles and topics, from expedition reports to scientific articles but all related to mountains.

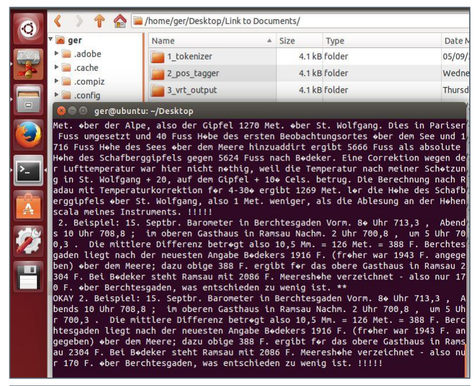
STRUCTURE CORRECTION

A particular challenge was the large number of volumes printed in German **Fraktur Font (46)**. Also books from the 1980s with sometime adventurous layouts provided problems for the system. Structure correction was done semi-manually with FEP, which adds markup to the text structure.



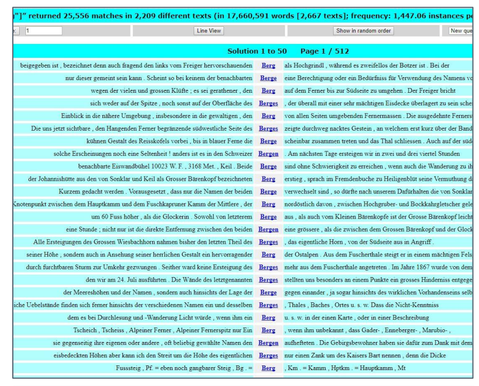
XML AND GIT-VERSIONING

From FEP an XML export of the text with structure annotations was taken as a corpus 1.0 and versionised under GIT. Departing from here a large number of post correction was done, especially on the *Fraktur*-related problems. The result is a relatively well corrected text to be sent into our POS-pipeline.



TAGGING AND ANNOTATION

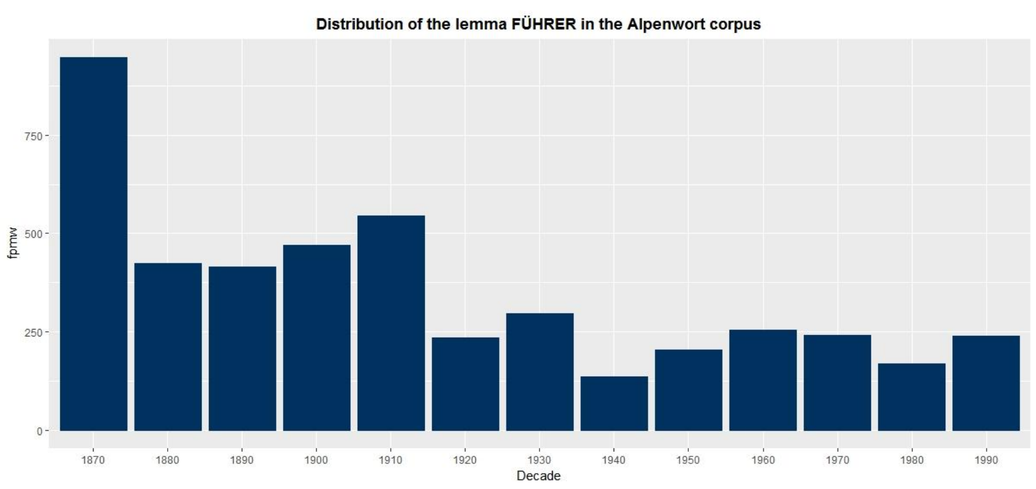
Our POS-pipe consists of TreeTagger modules which were trained on the closely related *text&berg corpus*. As part of the XML annotation we are working on enhancing the automatic classification and extraction of person names and geographical Names (NER) in a follow-up project.



POST CORRECTION & PUBLICATION

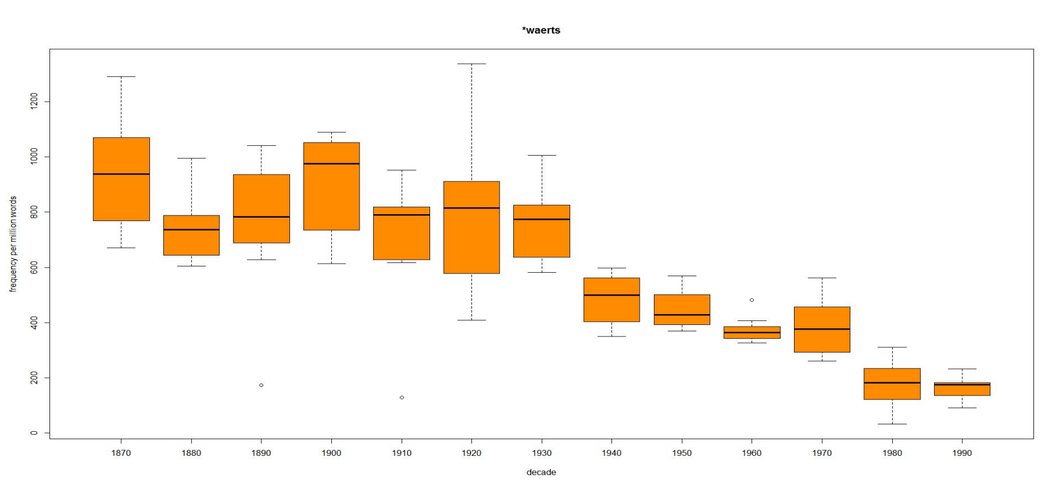
Tagging output fits the requirements of CWB's online CQPweb tool and is TEI conformant. The corpus is running on an in-house test-server. CQP frequency lists were used to do further post correction of OCR (with RegEx). The corpus also will available in XML-form as well as on the Hyperbase platform.

SAMPLE ANALYSIS



Left: distribution lemma deutsch

Right: derivations with -wärts



OUTLOOK

The project *Alpenwort* provides a fully POS-annotated thematic corpus of a very specific kind of functional literature. The corpus incorporates 123 German yearbooks and has a similar structure (markup and metadata) as its Swiss counterpart, the *text & berg* corpus (Bubenhofer et al. 2015). Both corpora are available on CQP-web and we will work with our Swiss colleagues to integrate the two corpora on a shared platform. As part of XML annotation we are working on the enhancement of POS-tagging and particularly on the automatic classification of named entities. The project *SEMOHI* (Semantics for Mountaineering History) aims to semantically enrich the *Alpenwort* corpus by identifying and tagging 1. places like mountains, regions, trails or huts, 2. people like mountaineers, guides or scientists, and 3. first ascent events like the first ascent of Großvenediger (by Josef Schwab in 1841) within the corpus. In order to achieve this goal specific gazetteers will be built from different data sources and connect them to the LOD (Linked Open Data) cloud. A further follow-up project to *Alpenwort* we are working on at the moment is *KEA*, in which we digitise the New Zealand Alpine Journal. We plan on a preliminary version of this corpus by the end of 2017.

References & Acknowledgements:
Bubenhofer, Noah , Martin Volk, Fabienne Leuenberger & Daniel Wüest (2015). Text+Berg-Korpus (Release 151_v01).
Evert, Stefan and Hardie, Andrew (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In Proceedings of the Corpus Linguistics 2011 Conference, University of Birmingham, UK.
Posch, Claudia and Gerhard Rampl (2017). Alpenwort. Korpus der Zeitschrift des Deutschen und Österreichischen Alpenvereins.
Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.
Vanni, Laurent and Damon Mayaffre (2013). Hyperbase Web Edition. Laboratoire BCL - UMR 7320, Université de Nice Sophia-Antipolis. http://hyperbase.unice.fr

poster download

ÖAW

universität innsbruck