



## The Project *Alpenwort* A corpus building bridges into different fields

Claudia Posch  
Gerhard Rampl

### *Alpenwort* - summiting strategy

‘The main goal of this proposal is to make this source accessible for the scientific community enhanced by exhaustive linguistic annotation.’





## Alpenwort – summing strategy

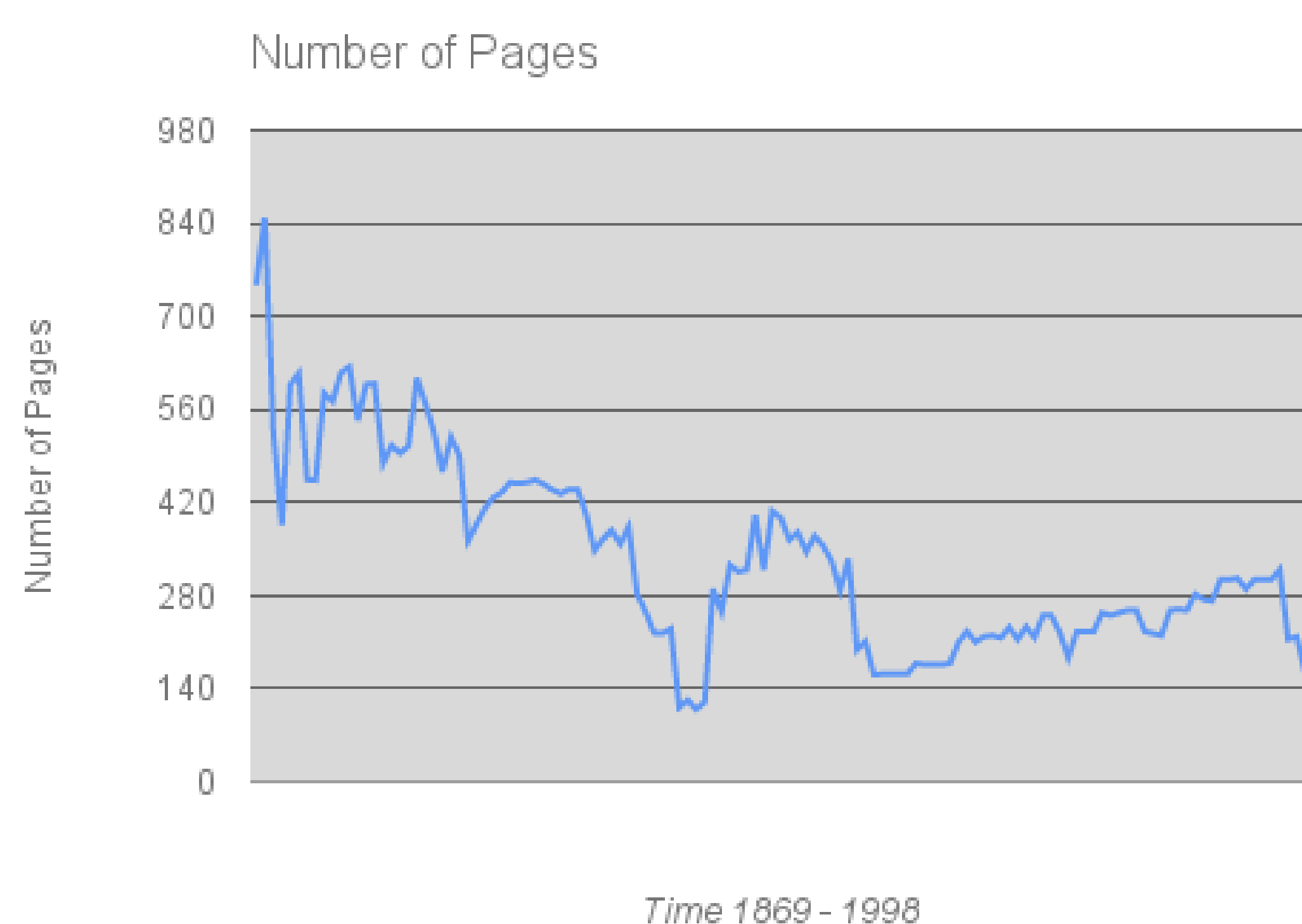
„The main goal of this proposal is to make this source accessible for the scientific community enhanced by exhaustive linguistic annotation.”

- » OCR and structure recognition (automated)
- » structure correction of OCR in FEP (semi-automated)
- » export of data in XML and version controlling (GITlab)
- » automated data annotation and enrichment with pipeline (automated)
- » pre-publication for testing (Hyperbase, CQP-Web)
- » OCR & text quality correction (semi-automatic)
- » re-tagging & publication

## Gear list

- » duration: 10/15/2014 – 10/14/2017
- » funding: go!digital call of the Austrian Academy of Sciences 2013
- » staff: Daniela Feistmantl, Bettina Larl, Irina Windhaber
- » partners:
  - Austrian Alpine Club,
  - DEA – Digitisation and Electronic Archiving (@UIBK),
  - Institute of Computational Linguistics of the University of Zurich,
  - DBIS – Databases and Information Systems (@UIBK),
  - Unit for Surveying and Geoinformation (@UIBK)
- » linguistically annotated corpus of the ZAV (126 years, 123 volumes)

## Gear list



min. 109 – max. 848 pages per volume

total: 43.383 pages

average: 338 pages

1915 – 1961 gothic script

all others in antiqua script

1961/1962 in both

## Cruxes climbed

- » the corpus is completed as planned
- » annotation and correction were done automatically
- » semi-automated annotation and correction have been added
- » meta-data TEI conformant:
  - one XML-file per volume
  - TEI-header with project information similar to BNC
  - annotation of volumes: file name, volume, year, decade
  - annotation of articles: article number, author, title, pages,...
- » POS-tagging: automatically done by pre-trained TreeTagger + enhanced lexicon (historical spellings!) → output in XML format/CQP-web format



## Cruxes climbed

- » automated structure recognition had to be corrected manually
- » particularly problematic with two-column pages
- » particularly problematic with fancy designs
- » every volume is different has a certain degree of individuality!
- » OCR-correction of Gothic Script (1916-1961)
- » use of corpus tools to find errors (AntConc)
- » partly systematic errors – semi-automated with RegEx
- » some paragraphs were toggled



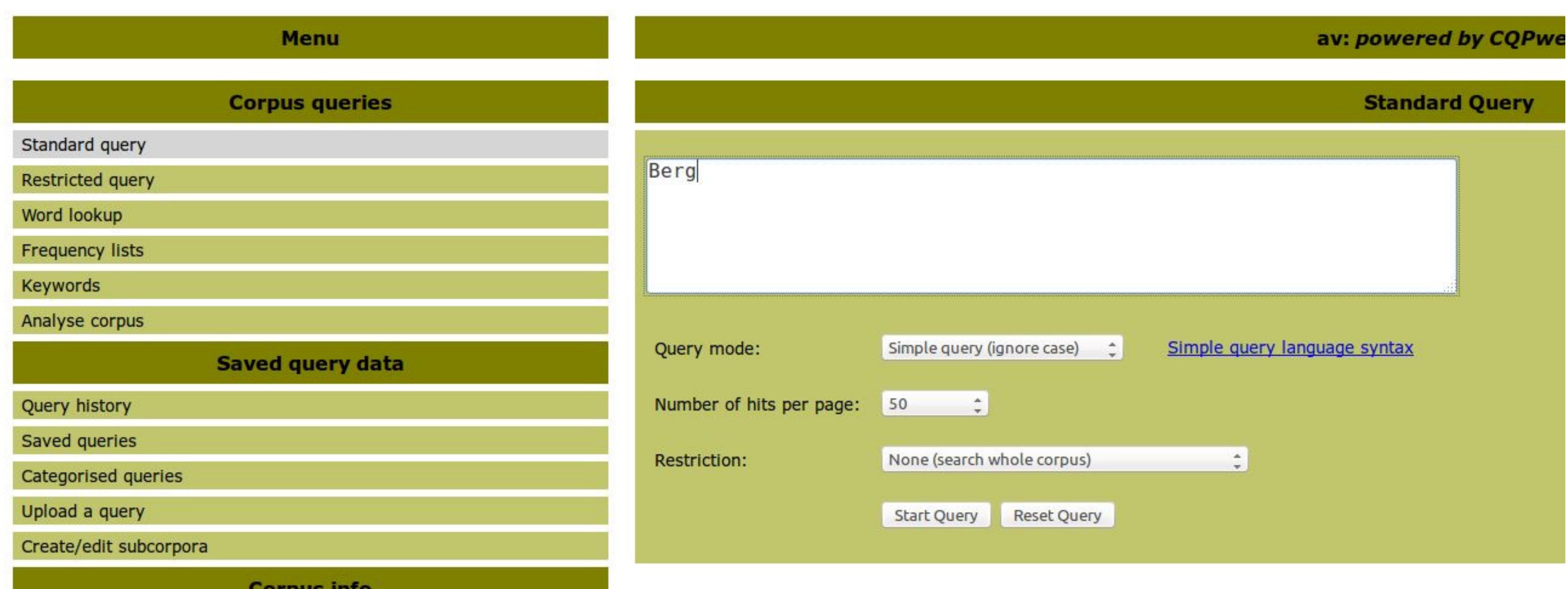
Bearbeitung eines Ackerfeldes im Pitzthale.

## Cruxes climbed

- » automated structure recognition had to be corrected manually
- » particularly problematic with two-column pages
- » particularly problematic with fancy designs
- » every volume is different has a certain degree of individuality!
- » OCR-correction of Gothic Script (1916-1961)
- » use of corpus tools to find errors (AntConc)
- » partly systematic errors – semi-automated with RegEx
- » some pages were toggled

## North face

- » implementation into *CQP-Web* on a local server at DBIS
- » implementation into *Hyperbase*

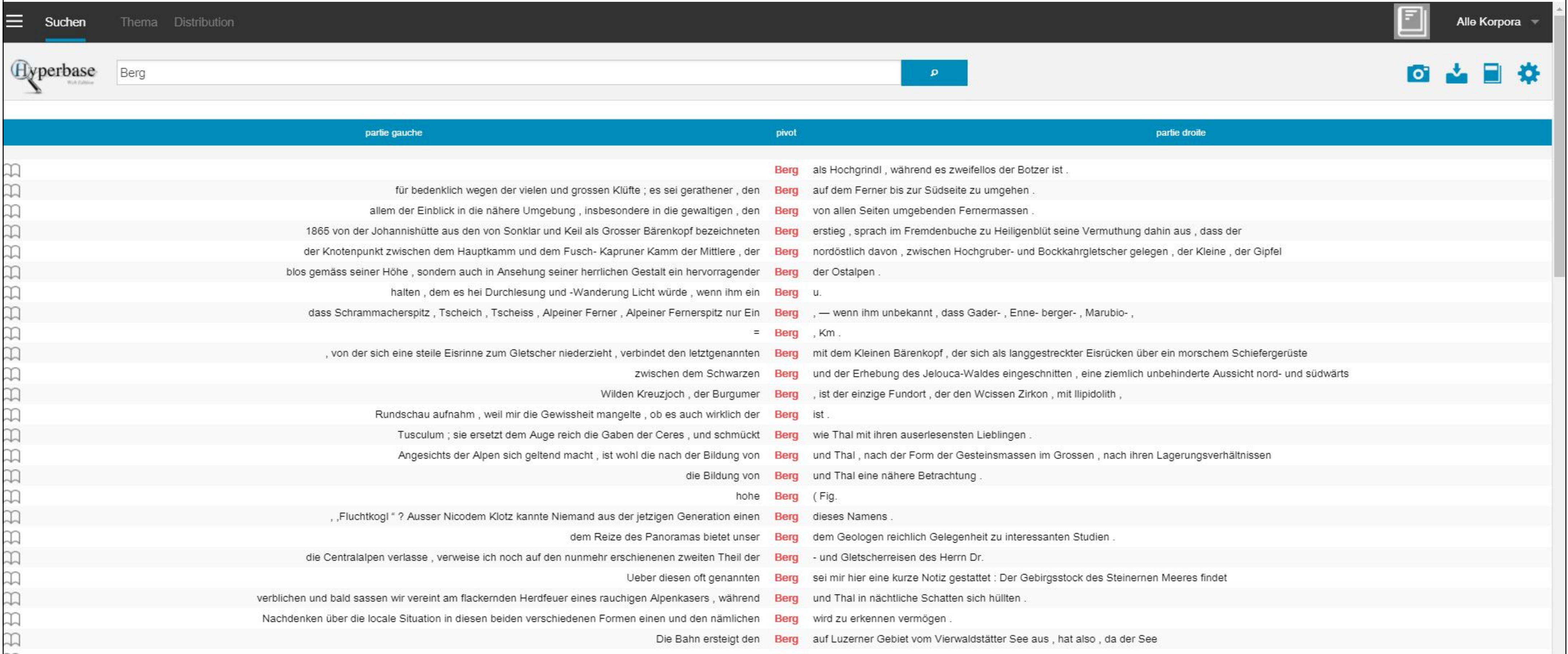


The screenshot displays the CQP-Web interface. On the left is a vertical menu with sections: 'Menu', 'Corpus queries' (containing 'Standard query', 'Restricted query', 'Word lookup', 'Frequency lists', 'Keywords', 'Analyse corpus'), 'Saved query data' (containing 'Query history', 'Saved queries', 'Categorised queries', 'Upload a query', 'Create/edit subcorpora'), and 'Corpus info'. The main area on the right is titled 'Standard Query' and features a large text input field containing 'Berg'. Below the input field are controls for 'Query mode' (set to 'Simple query (ignore case)' with a link to 'Simple query language syntax'), 'Number of hits per page' (set to 50), and 'Restriction' (set to 'None (search whole corpus)'). At the bottom of the main area are 'Start Query' and 'Reset Query' buttons.



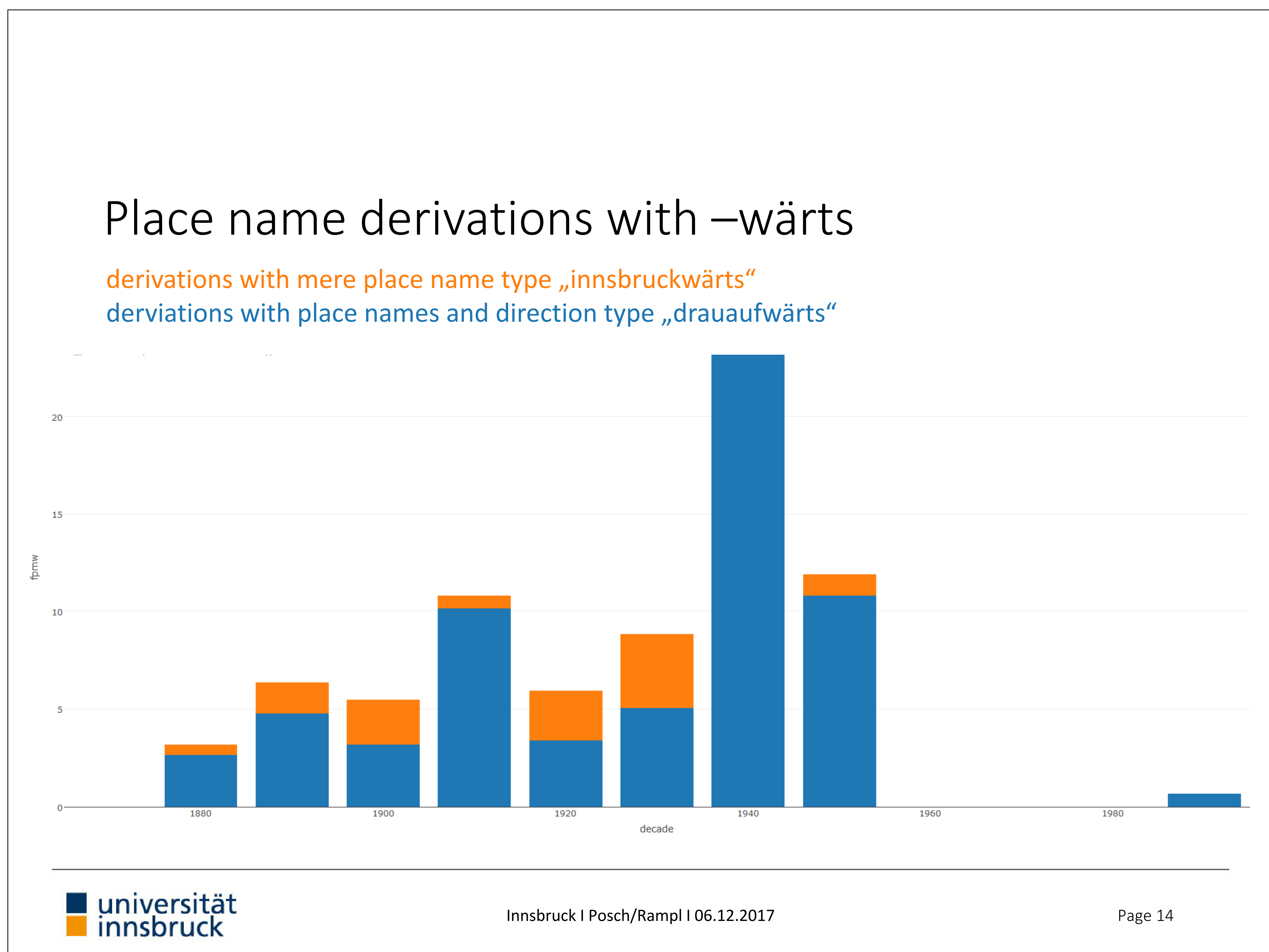
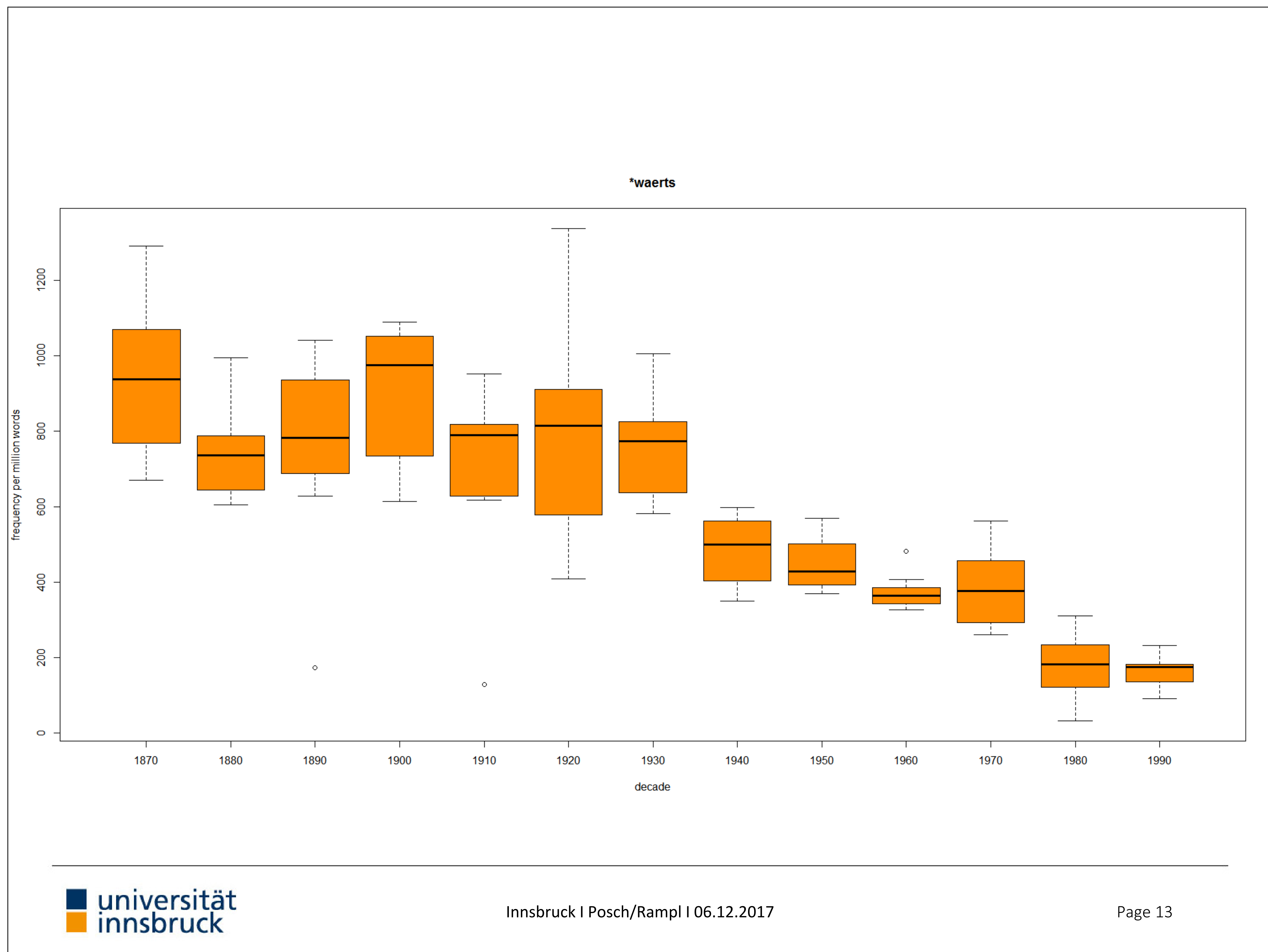
# South face

- » implementation into *CQP-Web* on a local server at DBIS
- » implementation into *Hyperbase*



# Bridges and knots

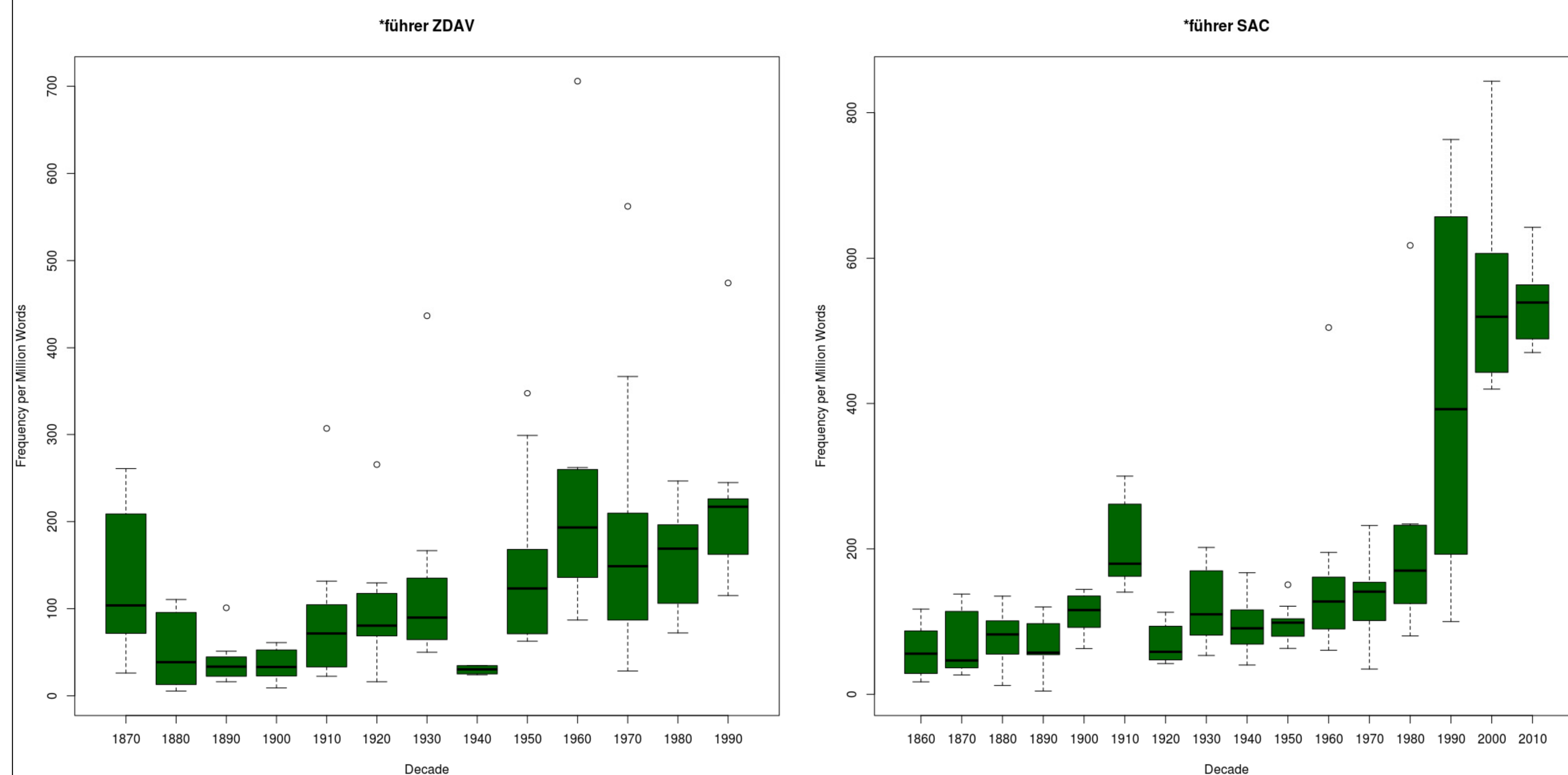
- » uses for other fields within linguistics
- » historical linguistics
- » discourse studies
- » onomastics



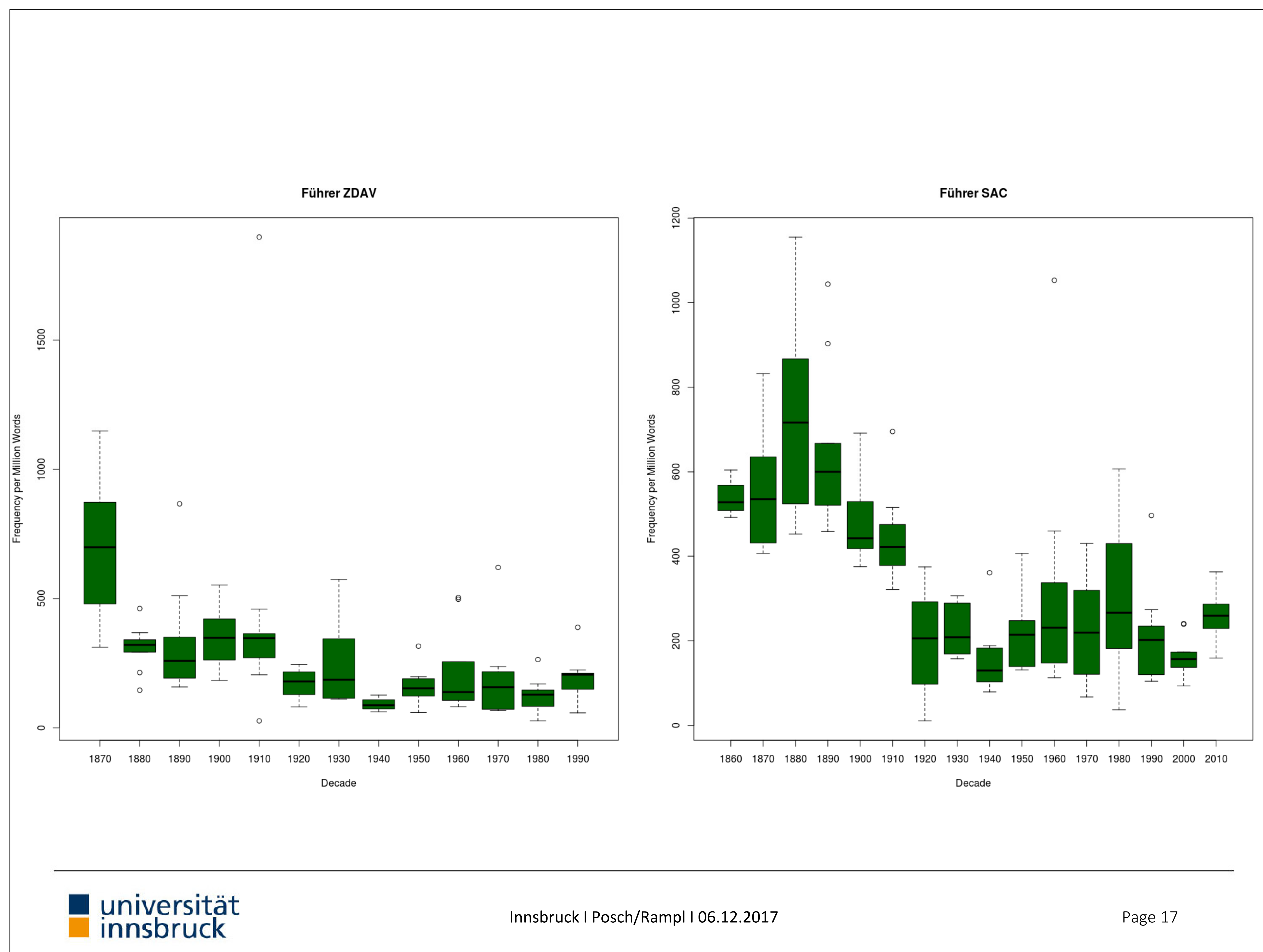


## Bridges and knots

- » Alpine Club
- » history of the Alpine Club (including world wars and Nazi-regime)
- » first ascents
- » exploration of unknown areas
- » re-publication of historic articles (Powder GUIDE)
- » search for people/places/events
- » links to other data possible (other corpora, archives, library, art depot, ...)







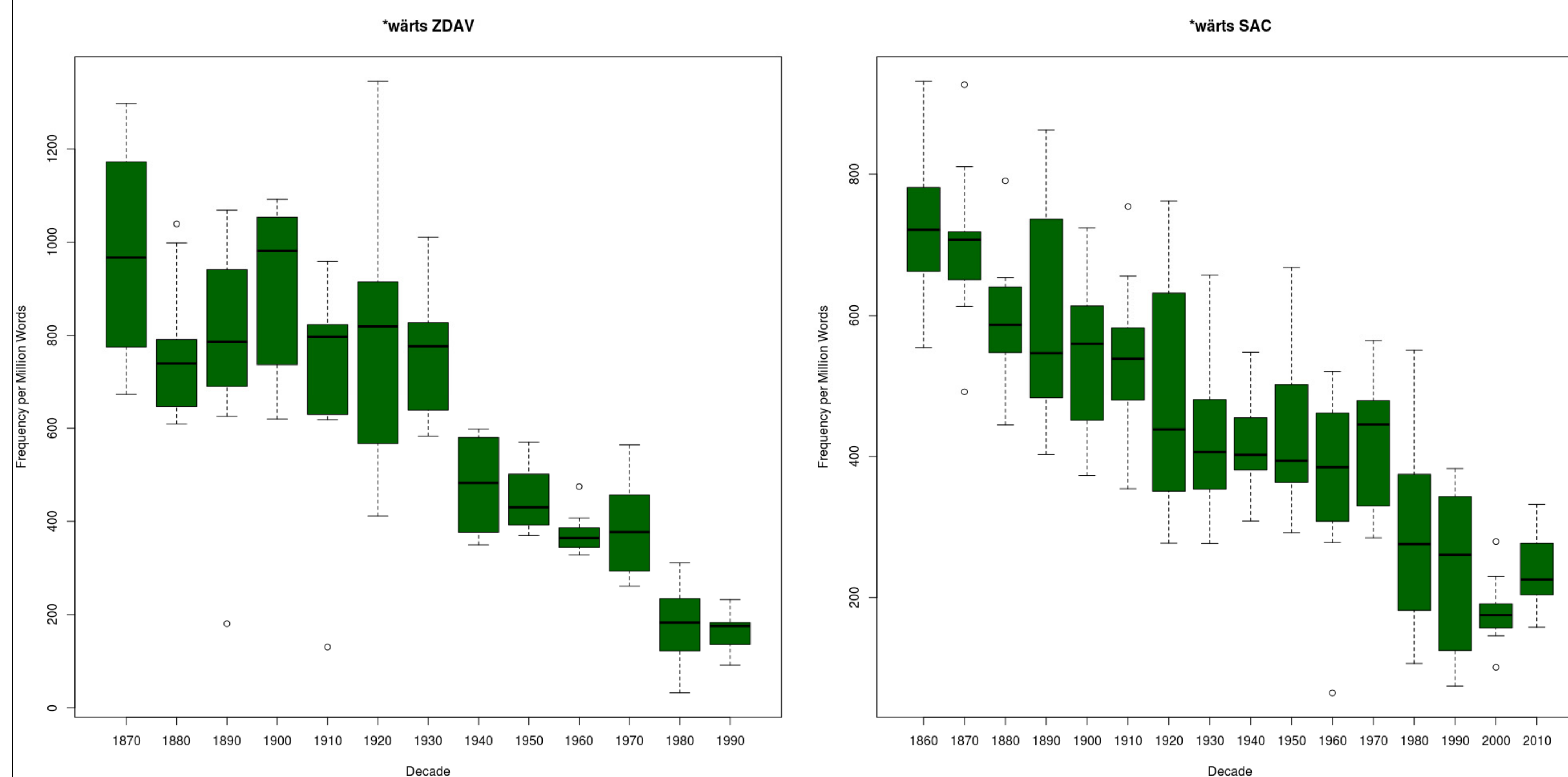
## Bridges and knots

- » Informatics and Geography
- » excellent training corpus for NER and language processing (including 19<sup>th</sup> century)
- » pipeline development for language processing
- » semantic tagging, geo-tagging
- » data linking to other data types (e.g. CIDOC, gazetteers)



## Panoramic view

- » corpus analyses
- » follow-up project SEMOHI – Semantics of Mountaineering History
- » follow-up project KEA (Keywords of the Alps)
- » comparison with existing corpora like text&berg
- » enrich annotation
- » link to other data (Lichtbilderverzeichnis, Autorenkatalog, biogr. Lexicon)
- » subsequent new corpora, e.g. Mitteilungen des Alpenvereins, sources from the Alpine archive, other languages
- » use of corpora in university teaching





## Panoramic view

- » publication of corpus on different platforms → end of 2017 (no pressure felt)
- » text-only XMLs (no POS) with TEI-conformant meta-data → on **zenodo**
- » publication of annotated and ready-to-search corpus on **Hyperbase**
- » information on release will be available via:
  - <http://sprawi.at>
  - on Twitter: @SpraWiTwit
  - newsletters
- » project website (under construction): <http://alpenwort.at>

