Text+Berg

Corpus of Alpine Texts from 1864 to 2014

Noah Bubenhofer
TU Dresden / University of Zurich

February 27, 2014
Fribourg

Noah Bubenhofer

Text+Berg Corpus

Goals

- Collect a Corpus of “Alpine” Texts
  - Mountaineering and travel reports: Routes, Achievements, Equipment, Contemplation,…
  - Popular science articles on mountain topics: Geology, Biology, Climatology, Tourism, Culture …
- in multiple languages
- for studies on language development, culture, zeitgeist, technology, …
- for geo-tagging, machine translation, …

Text+Berg Corpus

- Project in collaboration with Martin Volk and his group at UZH CL
- Current Corpus Release (1864 – 2011)
  - 147 years – 253 books (> 100,000 pages)
  - 90 mixed-language books 1864 – 1956
  - 110 parallel DE-FR books 1957 – 2011
  - 53 FR books (Echo des Alpes) 1872 – 1924

Language	Articles	Tokens
German	11'200	22.5 million
French	11'500	21.5 million
Italian	170	0.3 million
Romansch	16	0.05 million
Swiss-German	4	0.02 million

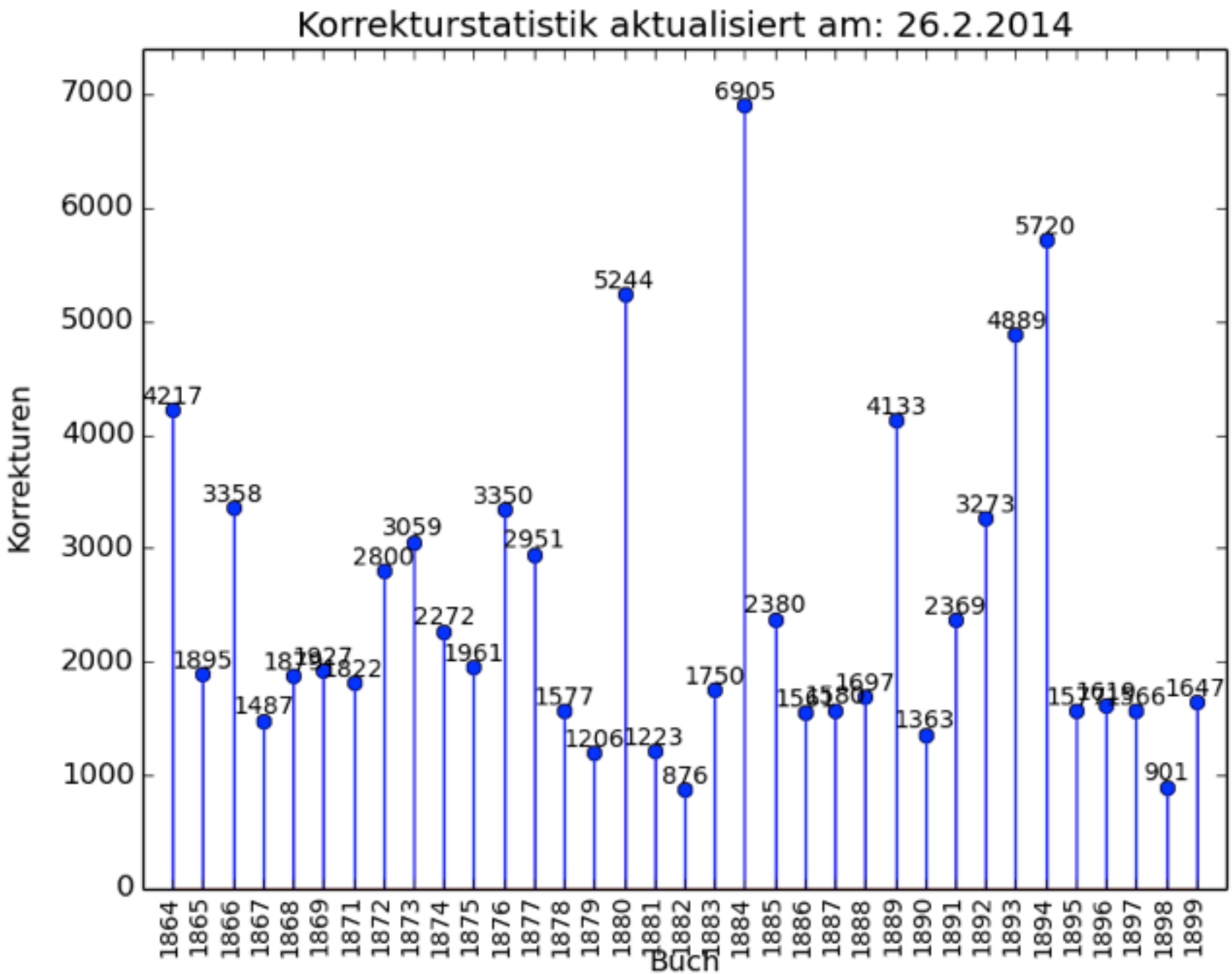
TECHNISCHE UNIVERSITÄT DRESDEN

University of Zurich[UZH]

Processing Pipeline

Digitizing Process

- OCR: Abbyy-Finereader and OmniPage
  - manual correction (crowdsourced) of 16 volumes
  - automated correction algorithms
    - common OCR errors (ii → ü etc.)
    - merging solutions of FineReader and OmniPage
- volumes 2001-2009: extraction from pdf
- volumes 2010-today: source is XML


Korrekturstatistik aktualisiert am: 26.2.2014
Korrekturen
Buch
0
1000
2000
3000
4000
5000
6000
7000
1864
1865
1866
1867
1868
1869
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
4217
1895
3358
1487
1822
3059
2800
2272
1961
3350
2951
1577
1206
5244
1223
876
1750
6905
2380
1697
4133
1363
2369
3273
4889
5720
1647
901

POS Tagging, Stemming

- Language identification (Lingua-Ident, Piotrowski)
- POS tagging:
  - TreeTagger with default libraries for DE, IT, EN
  - FR: TreeTagger library based on Le Monde treebank
  - Swiss German: German POS tags
  - Romansch: no POS tags
- Stemming: semi-automatic expansion of the lexicon

Named Entity Recognition

- Rule based approach
- Identification and Disambiguation of Geographical Entities: mountains, glaciers, lakes, cabins, towns
  - Gazetteer lookup: SwissTopo (156,755 names in 61 categories)
  - Gazetteer expansion: Common suffixes for mountain names (-horn, -stock, -grat)
- Identification of personal names

Named Entity Recognition

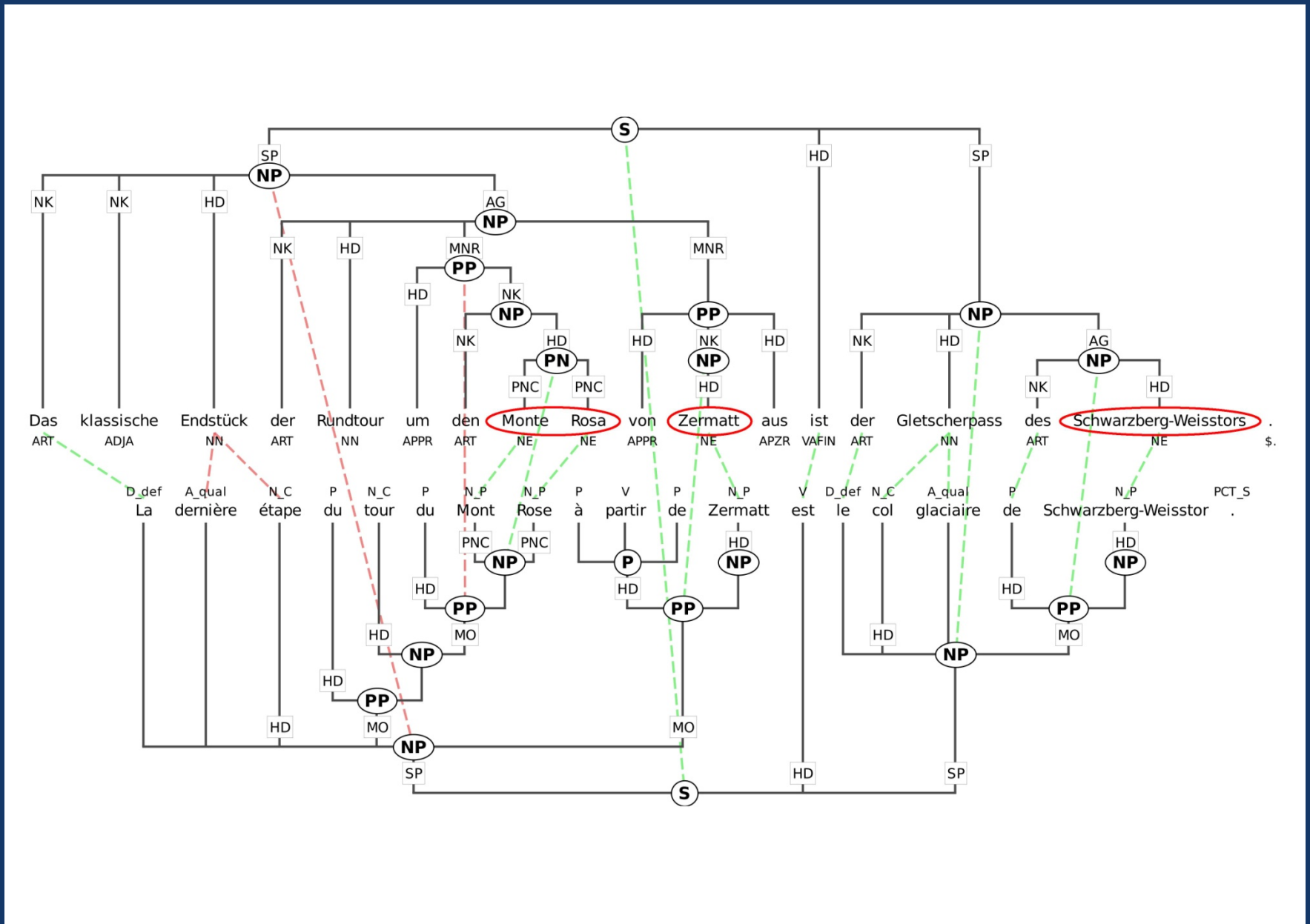
- 7 main mountains have ambiguous names (Dom, Esel, Jungfrau, Krone, Mönch, Ochse, Speer).
- 26 name types of minor mountains are nouns (the most frequent ones are Stock (10), Horn (6) and Stand (4)).
- curious ambiguous names
  - Bär, Löffel, Prosecco (field names in AG, FL, and TI)
  - Ast, Greuel, Gummi (the names of small villages or hamlets)

Distribution of the Data

- XML available for free for academic use
- Corpus Workbench, CQPweb-Interface

→ w

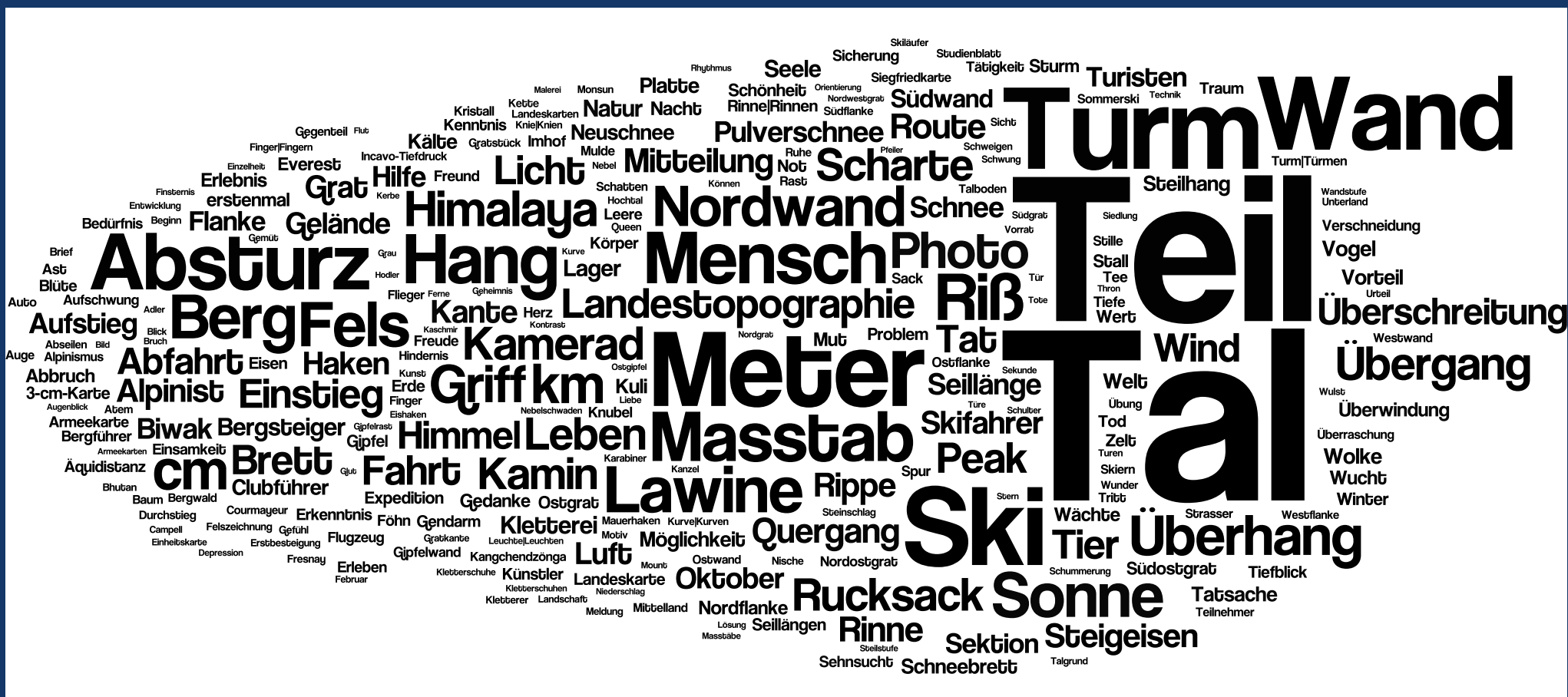
Parallel Treebank



Linguistic Analyses

- Example question:
  - Which words are significantly more prominent in one subcorpus rather than the other?
  - Here comparison: 1930-1949 vs. 1960-1979.

Corpus 1930-1949

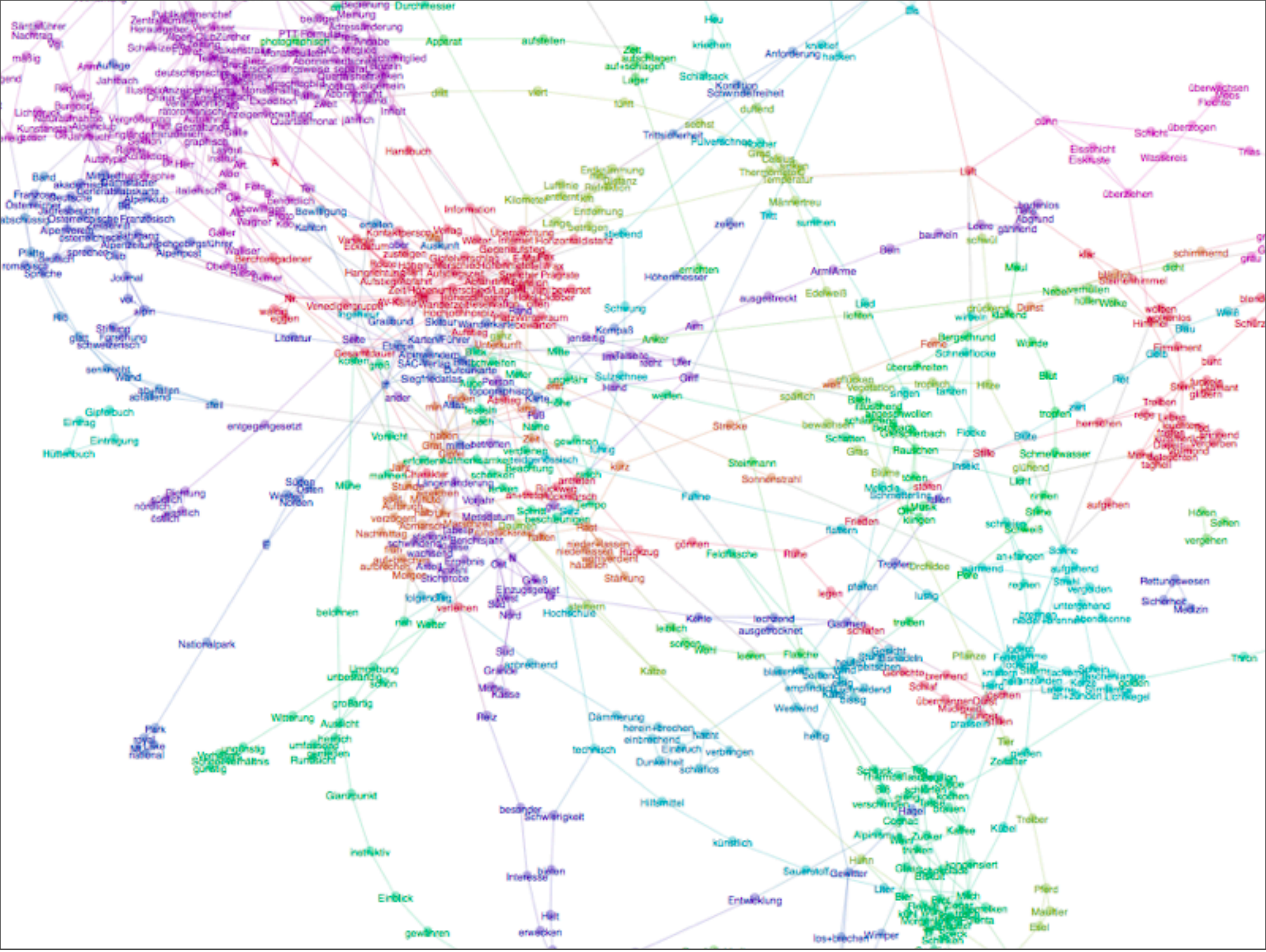


Noah Bubenhofer

Corpus 1960-1979

Noah Bubenhofer

Data Driven: Calculating Collocation Graphs



Samstag, 1. März 14

Calculating Linguistic Patterns

- n-grams:
  so verbringen wir
- complex n-grams:
  so [verb] [personal pronoun]
- calculating all possible combinations of complex n-grams for *n* = 3-5
- test of significance to get patterns prominent for specific time spans

Complex n-Grams (1880–1899)

- **ADV erreichten PPER ART NN**
  endlich erreichten wir den Aaresattel
  (Bald) nachher erreichten wir den Guggistafel
  Nun erreichten wir das Gebiet (des Kalkfelsens)
- **VVFIN APPR CARD Uhr**
  stiegen um 12 Uhr
  erreichten um 2 Uhr
  verließen um 3 Uhr
- **an der ADJA NN des**
  an der linken Seite des
  an der rechten Seite des
  an der anderen Seite des
  an der breiten Wand des

Textbeispiel 1888

"Wir hielten uns möglichst hoch auf der nördlichen, nach Süden fallenden Seite von Vaplona, gewannen das kleine Tobel, das gegen den Punkt 2547 m ansteigt, wateten durch dasselbe hinaus und erreichten den genannten Punkt um 8 Uhr 45 Min. Hier sahen wir etwa 200 m unter uns den Schottensee mit Schnee und Eis bedeckt. Schottenseefurke wäre vielleicht, der Wildseefurke (2515 m) entsprechend, der passendste Name für diese Scharte zwischen den Punkten 2647 m und 2650 m. Wir blieben nicht lange hier, sondern stiegen bald wieder links auf, theils über Schnee, theils über Verrucanotrümmer, und betraten den Gipfelpunkt 2650 m um 9 Uhr 15 Min. Es ging ein schwacher Windzug, und das Thermometer zeigte auf -20 C."

Jahrbuch 1888-1889: Eine Sectionsfahrt auf den Piz Sol (J. J. Schiesser)

Complex n-Grams (1930–1949)

- **ADV VVFIN ART ADJA NN .**
  Draussen erwachte ein neuer Tag .
  Dann kam ein trüber Tag .
  Nun naht das schwierigste Stück .
- **ADV VVFIN PPER VVINF**
  So lasst uns eilen
  jetzt heisst es handeln

Complex n-Grams (1930–1949)

- **dann VVFIN ART NN**
  dann VVFIN ART ADJA
  dann VVFIN PPER auf
  dann kündete die Gipfelglocke
  dann geschieht das Wunder
  dann folgt ein heikler
  dann standen wir auf (dem kühnen Gipfel)
- **KOUS PPER APPR ART NN VVFIN**
  KOUS PPER ART NN VVFIN
  Wie ich in den Riss einstieg
  als wir in der Gabel anlangten
  Bevor wir in das Couloir hinübersteigen
  während wir der Hütte zustrebten
  während wir die Steigeisen ablegten
  Als wir die Passhöhe erreichten

Textbeispiel 1932

"[N]ochmal wird angegriffen. Und endlich gelingt es mir, für die linke Hand ganz oben einen guten Griff zu schaffen. Die Rechte bohrt sich mit dem Eisbeil in der ersehnten Rampe ein Loch, der rechte Fuss steigt nochmal nach, und ich sehe über den Rand, indessen der Körper schwer nach aussen hängt. Sich ganz auf die rechte Hand verlassend, fährt die linke blitzschnell weit über den Rand in den Firn, ein Ruck, und ich liege verschnaufend auf dem Bauch, während die Füsse in der Luft baumeln."

Die Alpen 1932: Piz Bernina-Nordostflanke (Karl Schneider)

Noah Bubenhofer

Visualize Geocoded Data

Summary

- Building homogenious diachronic corpora:
  - NLP annotations
  - geocoding
  - metadata
- Using corpora for:
  - computational linguistics: adapting NLP techniques to domain specific and diachronic corpora; parallel corpora
  - linguistics, digital humanities: cultural studies, text linguistics, language history etc.

Danke – Thank you – Merci

Schweizer Alpen-Club SAC
Club Alpin Suisse
Club Alpino Svizzero
Club Alpin Svizzer

AUSTRIAN ACADEMY CORPUS

DOCUMENT TEXT INFORMATION DTI

Alpine CLUB

FNSNF
Schweizerischer Nationalfonds zur Förderung der wissenschaftlichen Forschung

Société de la Connaissance des Alpes

Universität Zürich[UZH]

ZENTRALBIBLIOTHEK ZÜRICH

Danke – Thank you – Merci

Etienne Ailloud
Adrian Althaus
Sara Bärtschi
Maya Bangerter
Adrian Beer
Andrea Bizetti
Alena Ciulla
Lenz Furrer
Pierpaolo Frasa
Anne Göhring
Reto Gubelmann
Alexandra Hartmann
Stefanie Herzig

Flavian Imlig
Silvan Jegen
Magdalena Jitca
Stéphanie Lehner
Torsten Marek
Nicole Michel
Stephanie Odok
Marcel Riedener
Beni Ruef
Patricia Scheurer
Rico Sennrich
Silvan Tarnutzer
Martin Warin

Literatur

Adler, Joseph. 2010. R in a Nutshell. 1. Aufl. O'Reilly.

Bätzing, Werner. 2003. Die Alpen: Geschichte und Zukunft einer europäischen Kulturlandschaft. C.H.Beck.

Belica, Cyril/Steyer, Kathrin. 2006. Korpusanalytische Zugänge zu sprachlichem Usus. In: AUC (Acta Universitatis Carolinae), Germanistica Pragensia. XX .

Bubenhofer, Noah. 2013. Quantitativ informierte qualitative Diskursanalyse. Korpuslinguistische Zugänge zu Einzeltexten und Serien. In: Roth, Kersten Sven/Spiegel, Carmen (Hrsg.) Angewandte Diskurslinguistik. Felder, Probleme, Perspektiven. Berlin: Akademie-Verlag (Diskursmuster - Discourse Patterns). 109–134.

Bubenhofer, Noah. 2009. Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. Berlin, New York: de Gruyter (Sprache und Wissen).

Bubenhofer, Noah/Scharloth, Joachim. 2013. Korpuslinguistische Diskursanalyse: Der Nutzen empirisch-quantitativer Verfahren. In: Warnke, Ingo/Meinhof, Ulrike/Reisigl, Martin (Hrsg.) Diskurslinguistik im Spannungsfeld von Deskription und Kritik. Berlin: Akademie-Verlag (Diskursmuster – Discourse Patterns). 147–168.

Bubenhofer, Noah/Scharloth, Joachim. 2011. Korpuspragmatische Analysen alpinistischer Literatur. In: Elmiger, Daniel/Kamber, Alain (Hrsg.) La linguistique de corpus – de l'analyse quantitative à l'interpretation qualitative / Korpuslinguistik – von der quantitativen Analyse zur qualitativen Interpretatino. Neuchâtel: Institut des sciences du langage et de la communication (Travaux neuchâtelois de linguistique). 241–259.

Bubenhofer, Noah/Schröter, Juliane. 2012. Die Alpen. Sprachgebrauchsgeschichte – Korpuslinguistik – Kulturanalye. In: Maitz, Péter (Hrsg.) Historische Sprachwissenschaft. Erkenntnisinteressen, Grundlagenprobleme, Desiderate. Berlin/Boston: de Gruyter (Studia Linguistica Germanica). 263–287.

Bubenhofer, Noah/Volk, Martin/Althaus, Adrian u. a. (Hrsg.). 2011. Text+Berg-Korpus (Release 145). Institut für Computerlinguistik, Universität Zürich.

Literatur

Ebling, Sarah/Sennrich, Rico/Klaper, David u. a.. 2011. Digging for names in the mountains: Combined person name recognition and reference resolution for German alpine texts. In: 5th Language & Technology Conference.

Evert, Stefan/The OCWB Development Team. 2010. The IMS Open Corpus Workbench (CWB) CQP Query Language Tutorial.

Feilke, Helmuth. 2000. Die pragmatische Wende in der Textlinguistik. In: Brinker, Klaus (Hrsg.) Text- und Gesprächslinguistik/Linguistics of Text and Conversation. Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science). 64–82.

Feilke, Helmuth/Linke, Angelika (Hrsg.). 2009. Oberfläche und Performanz. Untersuchungen zur Sprache als dynamische Gestalt. Berlin, New York: de Gruyter.

Gries, Stefan Thomas. 2008a. Dispersions and adjusted frequencies in corpora. In: International Journal of Corpus Linguistics. 13 . 403–437(35).

Gries, Stefan Thomas. 2009. Dispersions and adjusted frequencies in corpora: further explorations. In: Language and Computers. 71 (1). 197–212.

Gries, Stefan Thomas. 2008b. Statistik für Sprachwissenschaftler. Göttingen: Vandenhoeck & Ruprecht (Studienbücher zur Linguistik).

Gries, Stefan Thomas. 2010. Useful statistics for corpus linguistics. In: A mosaic of corpus linguistics Selected approaches. S. 269–291.

Gries, Stefan Thomas/Stefanowitsch, Anatol. 2004. Extending collostructional analysis. In: International Journal of Corpus Linguistics. 9 (1). 97–129.

Grupp, Peter. 2008. Faszination Berg die Geschichte des Alpinismus. Köln: Böhlau.

Literatur

Günther, Dagmar. 1998. Alpine Quergänge: Kulturgeschichte des bürgerlichen Alpinismus (1870 -1930). Campus Verlag.

Hermanns, Fritz. 1995. Sprachgeschichte als Mentalitätsgeschichte. Überlegungen zu Sinn und Form und Gegenstand historischer Semantik. In: Gardt, Andreas/Mattheier, Klaus/Reichmann, Oskar (Hrsg.) Sprachgeschichte des Neuhochdeutschen. Gegenstände, Methoden, Theorien. Tübingen: Niemeyer S. 69–101.

Jitca, Magdalena/Sennrich, Rico/Volk, Martin. 2011. From historic books to annotated XML: Building a large multilingual diachronic corpus. In: Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011. Universität Hamburg (Arbeiten zur Mehrsprachigkeit, Folge B. Working Papers in Multilingualism, Series B). 75–80.

Jorio, Marco. 2008. Geistiges Landesverteidigung. In: Historisches Lexikon der Schweiz (HLS).

Linke, Angelika. 1996. Sprachkultur und Bürgertum: zur Mentalitätsgeschichte des 19. Jahrhunderts. Stuttgart: Metzler.

Manning, Christopher D/Schütze, Hinrich. 2002. Foundations of Statistical Natural Language Processing. 5. Aufl. Cambridge, Massachusetts: The MIT Press.

Os, Charles van. 1989. Aspekte der Intensivierung im Deutschen. Tübingen: Narr (Studien zur deutschen Grammatik).

Perkuhn, Rainer/Belica, Cyril. 2006. Korpuslinguistik – Das unbekannte Wesen. Oder Mythen über Korpora und Korpuslinguistik. In: Sprachreport. 22 (1). 2–8.

Perkuhn, Rainer/Belica, Cyril/al-Wadi, Doris u. a.. 2005. Korpustechnologie am Institut für Deutsche Sprache. In: Schwitalla, Johannes/Wegstein, Werner (Hrsg.) Korpuslinguistik deutsch: synchron – diachron – kontrastiv. Würzburger Kolloquium 2003. Tübingen: Niemeyer S. 57–70.

Literatur

Scharloth, Joachim. 2005. Sprachnormen und Mentalitäten. Sprachbewusstseinsgeschichte in Deutschland im Zeitraum von 1766 bis 1785. Tübingen: Niemeyer (Reihe Germanistische Linguistik).

Scharloth, Joachim/Bubenhofer, Noah. 2011. Datengeleitete Korpuspragmatik: Korpusvergleich als Methode der Stilanalyse. In: Felder, Ekkehard/Müller, Marcus/Vogel, Friedemann (Hrsg.) Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen von Texten und Gesprächen. Berlin, New York: de Gruyter S. 195–230.

Scharloth, Joachim/Eugster, David/Bubenhofer, Noah (im Druck): Das Wuchern der Rhizome. Linguistische Diskursanalyse und Data-driven Turn. In: Busse, Dietrich/Teubert, Wolfgang (Hrsg.) Linguistische Diskursanalyse. Neue Perspektiven. Wiesbaden: VS Verlag.

Schiller, Anne/Teufel, Simone/Thielen, Christine. 1995. Guidelines für das Tagging deutscher Textcorpora mit STTS. Stuttgart.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees.

Sinclair, John. 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Steyer, Kathrin. 2004. Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven. In: Steyer, Kathrin (Hrsg.) Wortverbindungen – mehr oder weniger fest. Berlin, New York: de Gruyter (Institut für Deutsche Sprache. Jahrbuch 2003). 87–116.

Stremlow, Matthias. 1998. Die Alpen aus der Untersicht: von der Verheissung der nahen Fremde zur Sportarena: Kontinuität und Wandel von Alpenbildern seit 1700. Haupt.

Tognini-Bonelli, Elena. 2001. Corpus Linguistics at Work. Amsterdam: Benjamins (Studies in Corpus linguistics).

Literatur

Volk, Martin/Marek, Torsten/Sennrich, Rico (2010a): Reducing OCR errors by combining two OCR systems. In: ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010). S. 61–65.

Volk, Martin/Furrer, Lenz/Sennrich, Rico. 2011. Strategies for reducing and correcting OCR error. In: Sporleder, Caroline/Bosch, Antal van den/Zervanou, Kalliopi (Hrsg.) Language Technology for Cultural Heritage. Berlin: Springer (Theory and Applications of Natural Language Processing). 3–22.

Volk, Martin/Bubenhofer/Althaus, Adrian/Bangerter, Maya u. a. (2010b): Challenges in building a multilingual alpine heritage corpus. In: Seventh International Conference on Language Resources and Evaluation (LREC), Malta, 19 May 2010 – 21 May 2010.

Wall, Larry/Christiansen, Tom/Orwant, Jon. 2000. Programming Perl. O'Reilly.