

 **University of Zurich** UZH

 **Text+Berg digital**  
Projekt zur korpuslinguistischen Erschließung  
alpinistischer Literatur  
[www.textberg.ch](http://www.textberg.ch)

# Das Projekt Text+Berg

Martin Volk  
Institut für Computerlinguistik, Universität Zürich



**Text Mining**      **Sentiment Analysis**      **Machine Translation**

**e-Accessibility**      **Digital Humanities**

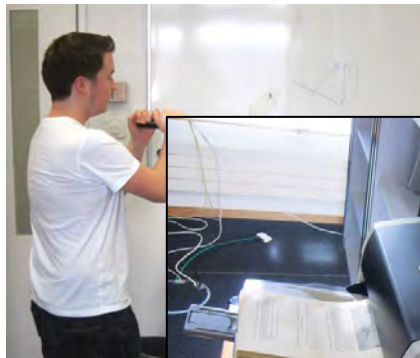
02.10.16 2


# Text+Berg Project



## 150 years of Alpinism in Switzerland:

- since 1864: “Jahrbuch des Schweizer Alpenclub”
- since 1925: New concept: „Die Alpen, Les Alpes, Le Alpi“
- today: monthly journal






**2 Drei Erstbesteigungen im Grimselgebiet:**  
**Klein-Lauteraarhorn (3742 m, höchste der 3 Spitzen), Nördliches Lauteraar-Rothorn (3478 m) und Vorder-Tierberg (3107 m).**

von  
**Paul Montandon** (Sektionen Bern und Bündnisalp)

**D**ie Herren Dr. Bob. v. Wyß, Albert Weber und ich stiegen am Nachmittag des 29. Juni 1904 von dem westlichen Grastriften der Oberaaralp auf den breiten oberen Hang, der sich unmittelbar den südlichen Abstürzen der ganzen Kette der Zinkenstöcke und Tierberge entlang zieht und damals größtenteils noch eine dicke Schneedecke trug. Die tiefe Grateinsenkung zwischen Hinter-Zinkenstock (3042 m) und Vorder-Tierberg (3107 m) erreichten wir mittelst Erklammerung einer kurzen, steilen Schlucht, die durch einen Block beinahe ausgefüllt ist, übrigens keine sonderlichen Schwierigkeiten bietet<sup>1)</sup>. Die Lücke liegt da, wo der Hauptgrat südlich vom Hinter-Zinkenstock im rechten Winkel nach Westen umbiegt, und es scheint für dieselbe der Name **Zinkenlöche** angezeigt. Sie ermöglicht einen guten Übergang vom mittleren Teil des Oberaar-gletschers nach dem Pavillon Dollfus. Auch der Hinter-Zinkenstock, das erste, kurze Stück vielleicht ausgenommen, ist von hier aus ganz leicht zu ersteigen. — Bis zu dieser Stelle, der Zinkenlöche, waren wir in einem Tempo vorgerückt, das uns jedenfalls den Rekord der Langsamkeit hätte eintragen müssen. Nun aber, es war 4 Uhr 40 Min.

1) Wir schlagen vor, den Punkt 3107 der Stieghilfskarte „Vorder-Tierberg“ zu nennen und den „Tiersberg“ (1292 m) in „Hinter-Tierberg“ umzutauften. Es scheint uns dies rationeller, als für Punkt 3107 denselben Namen Tyndallhorn zu adoptieren, den die meisten Bergsteiger, Alpine Journal, Schweizer Anzeiger, pag. 353 vorschlagen.



**Drei Erstbesteigungen im Grimselgebiet:**  
**Klein-Lauteraarhorn (3742 m, höchste der 3 Spitzen), Nördliches Lauteraar-Rothorn (3478 m) und Vorder-Tierberg (3107 m).**

von  
**Paul Montandon** (Sektionen Bern und Bündnisalp)

**D**ie Herren Dr. Bob. v. Wyß, Albert Weber und ich stiegen am Nachmittag des 29. Juni 1904 von dem westlichen Grastriften der Oberaaralp auf den breiten oberen Hang, der sich unmittelbar den südlichen Abstürzen der ganzen Kette der Zinkenstöcke und Tierberge entlang zieht und damals größtenteils noch eine dicke Schneedecke trug. Die tiefe Grateinsenkung zwischen Hinter-Zinkenstock (3042 m) und Vorder-Tierberg (3107 m) erreichten wir mittelst Erklammerung einer kurzen, steilen Schlucht, die durch einen Block beinahe ausgefüllt ist, übrigens keine sonderlichen Schwierigkeiten bietet<sup>1)</sup>. Die Lücke liegt da, wo der Hauptgrat südlich vom Hinter-Zinkenstock im rechten Winkel nach Westen umbiegt, und es scheint für dieselbe der Name **Zinkenlöche** angezeigt. Sie ermöglicht einen guten Übergang vom mittleren Teil des Oberaar-gletschers nach dem Pavillon Dollfus. Auch der Hinter-Zinkenstock, das erste, kurze Stück vielleicht ausgenommen, ist von hier aus ganz leicht zu ersteigen. — Bis zu dieser Stelle, der Zinkenlöche, waren wir in einem Tempo vorgerückt, das uns jedenfalls den Rekord der Langsamkeit hätte eintragen müssen. Nun aber, es war 4 Uhr 40 Min.

1) Wir schlagen vor, den Punkt 3107 der Stieghilfskarte „Vorder-Tierberg“ zu nennen und den „Tiersberg“ (1292 m) in „Hinter-Tierberg“ umzutauften. Es scheint uns dies rationeller, als für Punkt 3107 denselben Namen Tyndallhorn zu adoptieren, den die meisten Bergsteiger, Alpine Journal, Schweizer Anzeiger, pag. 353 vorschlagen.

<span class=font5> &gt;jie Herren <i> Dr. Bob. v. Wyß, Albert Weber </i> und ich stiegen am Nachmittag des 29. Juni 1904

&gt;jie  
Herren  
Dr.  
Bob.  
v.  
Wyß,  
Albert  
Weber  
und  
ich  
stiegen  
am  
Nachmittag  
des  
29.  
Juni  
1904

<s n="6-5" lang="de">

&gt;jie  
Herren  
Dr.  
Bob.  
v.  
Wyß  
,  
Albert  
Weber  
und  
ich  
stiegen  
am  
Nachmittag  
des  
29.  
Juni  
1904



```

<s n="6-5" lang="de">
  <w n="6-5-1" >&gt;jie</w>
  <w n="6-5-2" >Herren</w>
  <w n="6-5-3" >Dr.</w>
  <w n="6-5-4" >Bob.</w>
  <w n="6-5-5" >v.</w>
  <w n="6-5-6" >Wyß</w>
  <w n="6-5-7" >,</w>
  <w n="6-5-8" >Albert</w>
  <w n="6-5-9" >Weber</w>
  <w n="6-5-10" >und</w>
  <w n="6-5-11" >ich</w>
  <w n="6-5-12" >stiegen</w>
  <w n="6-5-13" >am</w>
  <w n="6-5-14" >Nachmittag</w>
  <w n="6-5-15" >des</w>
  <w n="6-5-16" >29.</w>
  <w n="6-5-17" >Juni</w>
  <w n="6-5-18" >1904</w>

```

```

<s n="6-5" lang="de">
  <w n="6-5-1" pos="ADJA" lemma="unk">&gt;jie</w>
  <w n="6-5-2" pos="NN" lemma="Herr">Herren</w>
  <w n="6-5-3" pos="NN" lemma="Dr.">Dr.</w>
  <w n="6-5-4" pos="NE" lemma="unk">Bob.</w>
  <w n="6-5-5" pos="APPR" lemma="von">v.</w>
  <w n="6-5-6" pos="NE" lemma="unk">Wyß</w>
  <w n="6-5-7" pos=",$" lemma=",">,</w>
  <w n="6-5-8" pos="NE" lemma="Albert">Albert</w>
  <w n="6-5-9" pos="NE" lemma="Weber">Weber</w>
  <w n="6-5-10" pos="KON" lemma="und">und</w>
  <w n="6-5-11" pos="PPER" lemma="ich">ich</w>
  <w n="6-5-12" pos="VFIN" lemma="steigen">stiegen</w>
  <w n="6-5-13" pos="APPRART" lemma="am">am</w>
  <w n="6-5-14" pos="NN" lemma="Nachmittag">Nachmittag</w>
  <w n="6-5-15" pos="ART" lemma="d">des</w>
  <w n="6-5-16" pos="ADJA" lemma="29">29.</w>
  <w n="6-5-17" pos="NN" lemma="Jun">Juni</w>
  <w n="6-5-18" pos="CARD" lemma="@card@">1904</w>

```

## Text+Berg Corpus



- Current Corpus Release (1864 – 2014)
- 151 years – 262 books (> 100,000 pages)
  - 90 mixed-language books 1864 – 1956
  - 53 FR books (*Echo des Alpes*) 1872 – 1924
  - 119 parallel DE-FR(-IT) books 1957 – 2014

Language	Articles	Tokens
German	12'180	23.5 million
French	12'930	23.0 million
Italian	1190	0.9 million
Romansch	18	50,000
Swiss-German	3	20,000

11

## Text+Berg Corpus: Types

= different words per language in

- Jahrbuch des SAC (1864 – 1923)
- Die Alpen (1925 – 2014)

Language	Tokens	Word Form Types	Word Form Types (lower)	Lemma Types	unknown Lemmas
German	22.81 million	763,000	722,000	294,000	790,000
French	14.64 million	290,000	266,000	62,000	565,000

## Reasons for high type counts

- old and new spellings
  - *Mittheilungen* vs. *Mitteilungen*
- standard and dialect spellings
  - *Weisshorn* vs. *Wysshorn*
- the compounding language German
  - *Hauptstationen der berneroberländischen Alphornbläser* (Die ALPEN 1925)
- OCR errors

02.10.16

13

## Reasons for high type counts

mountaineering terminology

- *Seil*
  - + *Gletscherseil* + *Reserveseil* + *Notseil* + *Dreißigmeterseil* + *12-Millimeterseil* + ...
- *Reissverschluss*
  - + *Gamaschen-Reissverschluss*
  - + *Seitenventilations-Reissverschluss*
  - + *2-Wege-Front-Reissverschluss*
  - + *Klemmschutz-Reissverschluss*
  - + *Armreißverschluß* ...

02.10.16

14

## Challenge: OCR Accuracy

Positive 😊

- No Gothic font

~~ten des Stromes, wenn ein Schiff umschlug oder zerfahelte. Durch~~

- Antiqua since the start in 1864

<b>I. Chronik des Club.</b> Von <i>A. R.</i> . . . . .	1.
<b>II. Fahrten im Clubgebiet.</b> . . . . .	15.
1) Generalbericht über die Excursionen im officiellen Gebiete während des Sommers 1863. Von Dr. <i>Th.</i> <i>Simler.</i> . . . . .	17.

## Challenge: OCR Accuracy

Problematic 😞

- German spelling variants
  - 1800s: *Nachtheil, passiren, successive*
  - 1900s: *Nachteil, passieren, sukzessive*
- Swiss German
  - 1880: *keiner macht Anstalten, seine Sachen wieder einzupacken. „Ja was isch jetze, wei mer eigentlich ufe?“ Peter schaut sinnend hinauf ...*



# Correcting OCR Errors

- Use an external lexicon
  1. Classify each corpus word as known / unknown
    - + *Eigerbesteigungen*
    - + *Erstüberquerung*
    - + *Selbstmordlandschaft*
    - *Hauptsfrukturlinien*
    - *Männergesaugvereins*
  2. For each unknown word, find its closest variant in the corpus.
    - *Hauptsfrukturlinien* → *Hauptstrukturlinien*
    - *Männergesaugvereins* → *Männergesangvereins*

## Digitale SAC-Jahrbücher korrigieren

Das «Text+Berg»-Projekt des Instituts für Computerlinguistik an der Universität Zürich widmet sich der Digitalisierung und Aufbereitung der SAC-Jahrbücher und der Zeitschrift «Die Alpen» von 1864 bis heute. Bei der automatisierten Textdigitalisierung geschehen leider unvermeidbare Fehler, die nicht automatisch behoben werden können, sondern einer manuellen Korrektur bedürfen. Mit der Internetplattform SAC-Kokos haben Sie jetzt die Möglichkeit, den Wissenschaftlern bei der Korrektur von falsch erkannten Buchstaben und Wörtern zu helfen, und erhalten zugleich einen spannenden Einblick in die SAC-Jahrbücher des 19. Jahrhunderts. So zum Beispiel in den Artikel über die «Erste Besteigung der Surettahörner» (1869) oder die Ausführungen zu «photographischen Aufnahmen im Hochgebirge» (1890). Weitere Informationen finden Sie unter: <http://kokos.cl.uzh.ch>.

Martin Volk, Projektleiter «Text+Berg»,  
Institut für Computerlinguistik Zürich



Wurde im SAC unangenehm Touristen belästigt, fällt nicht unter das Gesetz für Privatreisende des Kantons Uri. Foto: Bruno Huser

**Logistikpreis für SAC-Hütten**

Der SAC ist Träger des Swiss Logistics Public Award. Er erhält den Preis «Für die logistische Hochleistung der Ver- und Entsorgung der SAC-Hütten», wie GSI Schweiz, der Fachverband für nachhaltigen Wirtschaftsgüterverkehr, mittelst. Ziel des Verbands ist die ganzheitliche Optimierung von Waren- und Informationsflüssen. Der Preis wird seit 18 Jahren vergeben. Preiszähler waren unter anderem das Paléo Festival in Nyon und die Schweizer Wanderwege.

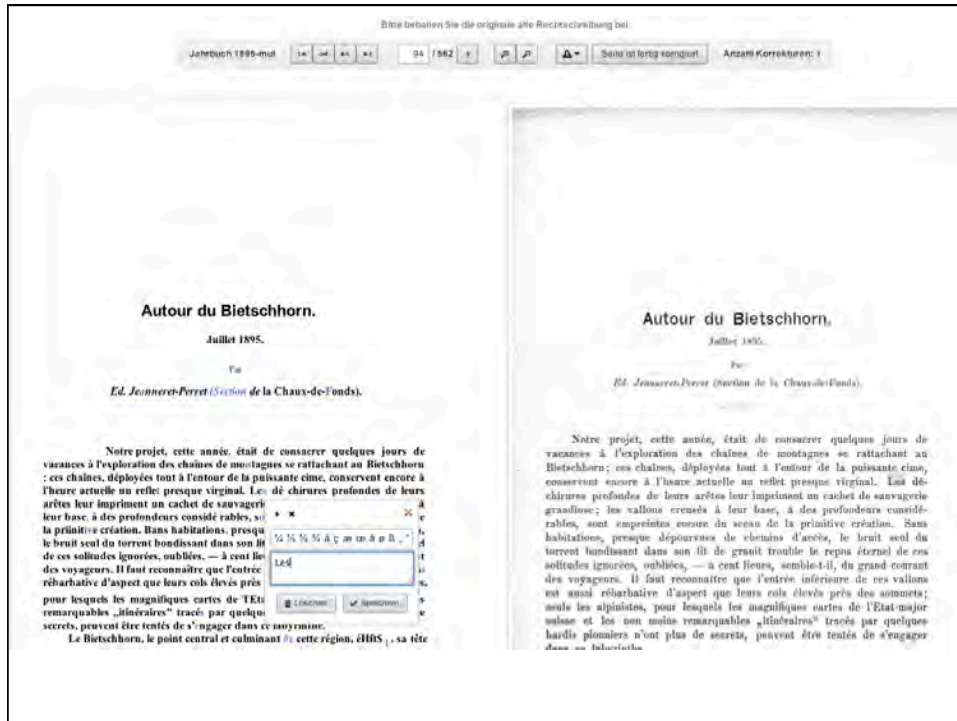
**Bundesgeoportal: neuer Kartenviewer**

Das Geoportal des Bundes wurde beim «Euro Cloud Award 2014» ausgezeichnet. Damit werden innovative Cloud-Lösungen geehrt, welche sich durch Originalität, Innovation, Kreativität und Effizienz auszeichnen, wie es in einer Mitteilung heisst. Der Kartenviewer [mobile.map.geo.admin.ch](http://mobile.map.geo.admin.ch) bringt in seiner neuen Version verbesserte Unterstützung für Tablets und Smartphones. Neue Kartenkategorien vereinfachen den Zugang zu den Geodaten des Bundes.

**Der neue Kartenviewer des Bundes**

Auf dieser Karte ist der Niederschlagshaushalt zu sehen. Braun ist sehr trocken, Blau sehr feucht.





SACKOKOS

Kokos | FAQ | SAC Jahrbücher

Eingeloggt als mvolk Abmelden

Universität Zürich

Bitte behalten Sie die originale alte Rechtschreibung bei.

Jahrbuch 1895-mit 94 / 562 Seite ist fertig korrigiert Anzahl Korrekturen: 1

#1: OCR correction is part of a reading experience

- »Correct errors while your are reading a page of interest.«
- Choose by article, search word, ...

### Autour du Bietschhorn.

Juillet 1895.

Par  
Ed. Jeanneret-Perret (Section de la Chaux-de-Fonds).

Notre projet, cette année, était de consacrer quelques jours de vacances à l'exploration des chaînes de montagnes se rattachant au Bietschhorn : ces chaînes, déployées tout à l'entour de la puissante cime, conservent encore à l'heure actuelle un relief presque virginal. Les déchirures profondes de leurs arêtes leur impriment un cachet de sauvagerie grandiose; les vallons creusés à leur base, à des profondeurs considérables, sont empreintes encore du sceau de la primitive création. Sans habitations, presque dépourvues de chemins d'accès, le bruit seul du torrent bondissant dans son lit de granit trouble le repos éternel de ces solitudes ignorées, oubliées, — à cent lieues, semble-t-il, du grand courant des voyageurs. Il faut reconnaître que l'entrée inférieure de ces vallons est aussi rébarbative d'aspect que leurs cols élevés près des sommets; seuls les alpinistes, pour lesquels les magnifiques cartes de l'Etat-major suisse et les non moins remarquables itinéraires tracés par quelques hardis pionniers n'ont plus de secrets, peuvent être tentés de s'engager dans ces montagnes.

- #4: Keep them motivated...
- Light-weight gamification by ranking statistics
  - Can be strongly motivating for some...

**Ihre persönliche Korrekturstatistik**

Rang	Benutzername	Korrekturen
1	anonymisiert	81050
2	anonymisiert	46987
3	anonymisiert	36226

**Ihr Rang**

<b>6</b>	<b>SimonClematide</b>	<b>13726</b>
----------	-----------------------	--------------

**Weitere Statistiken**

Total bisherige Korrekturen	256410
Durchschnittliche Korrekturen pro Benutzer	3611.0

Hinweis: Alle Statistiken werden nur einmal pro Tag erneuert.

- #5: Motivate the work on uncorrected pages
- Trigger the motivation to tidy up
  - Not all pages are fun to read (club news, scientific reports on plants, geology, glaciers)

**Übersicht der fertig korrigierten Seiten: Jahrbuch 1873**

## Corpus Usage

- Records and Observations

## Records: The longest word

*Gesundheitswiederherstellungsmittelmischungsverhältniskundiger* (62 letters)

*Unterwegs trafen wir noch einen befreundeten Hirschjäger, der in "Zivil" als "**Gesundheitswiederherstellungsmittelmischungsverhältniskundiger**" wirkt und nebenbei beim Schießen und Kegelschieben als Champion auftritt.*

In: Matth. Thöny (1905) "Ein Besuch der Sulzfluhhöhlen"

## Records: The longest word

- without quotation marks
  - *sechstausendsechshundertfünfundfünfzig* (38 letters, 1948)
  - *Haftpflichtversicherungsgesellschaft* (36 letters, 1914)
  - *Schwindelminenaktiengesellschaften* (34 letters, 1912)
- with hyphens
  - *«Inestäche-umeschloh-durezieh-ond-abeloh»-Technik* (49 chars, 2006)
  - *MIGROS-Sporthilfe-Nachwuchsförderungsprojekts* (45 chars, 2000)
  - *Lawinenverschüttetensuchgerät-Übungsgelände* (43 chars, 2008)

## Lemmatisation issues

### Verbs with separated prefixes

- *fängt der Steinschlag an*
  - `<w pos="VVFİN" lemma="an+fängen">fängt</w>`
  - `<w pos="ART" lemma="d">der</w>`
  - `<w pos="NN" lemma="Steinschlag">Steinschlag</w>`
  - `<w pos="PTKVZ" lemma="an">an</w>`
- *das nahe Ziel spornt uns an*
- *das schlechte Wetter hält an*
- *Ich schreibe es den Sturmwinden zu*

## Records: The longest span

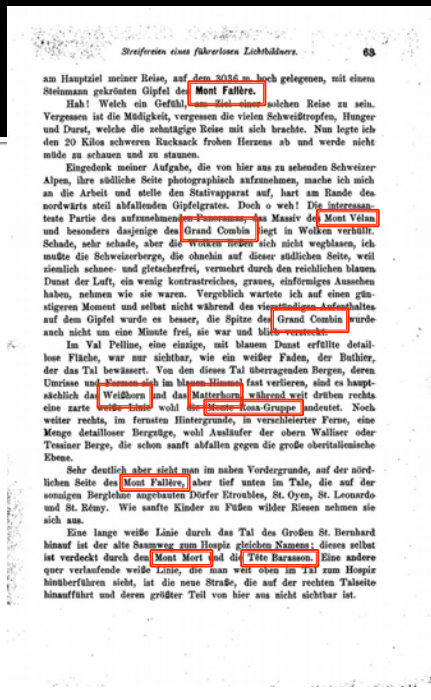
The longest span from German verb to separated prefix

- 57 words between **stürzten + ab**
- *Auf dem Hochfirn der Jungfrau stürzten der 24jährige stud. jur. Emil Frick und sein 20jähriger Bruder Paul, stud. chem., aus Zürich, welche vom Rottal aus die Besteigung unternommen hatten und durch ungünstiges Wetter zu einem Biwak, zusammen mit einer andern Partie, unter dem Gipfel genötigt worden waren, beim Fortsetzen der Tour, nachdem sie sich von der andern Partie getrennt hatten, etwa 60m tief ab.*

Aus: "Alpine Unglücksfälle 1914 und 1915" (Jahrbuch des SAC, 1914)

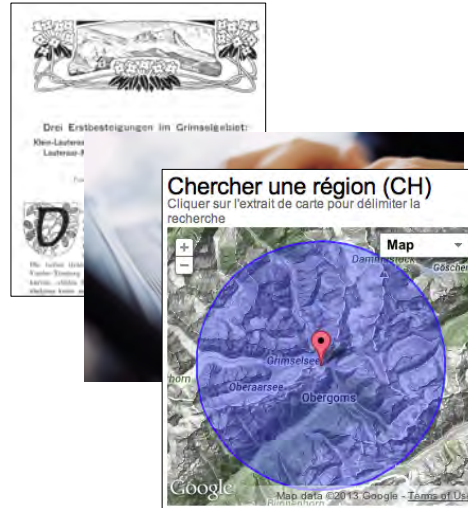
## Challenge: Geo-Tagging

- Identification and Disambiguation of Geographical Entities
  - Mountains
  - Cabins
  - Glaciers
  - Lakes
  - Cities / Towns
  - Valleys



## Summary

- How to preserve heritage documents in electronic form?
- How to make heritage documents widely accessible?
- How to do new things with old texts?



## Danke – Merci – Thank you

