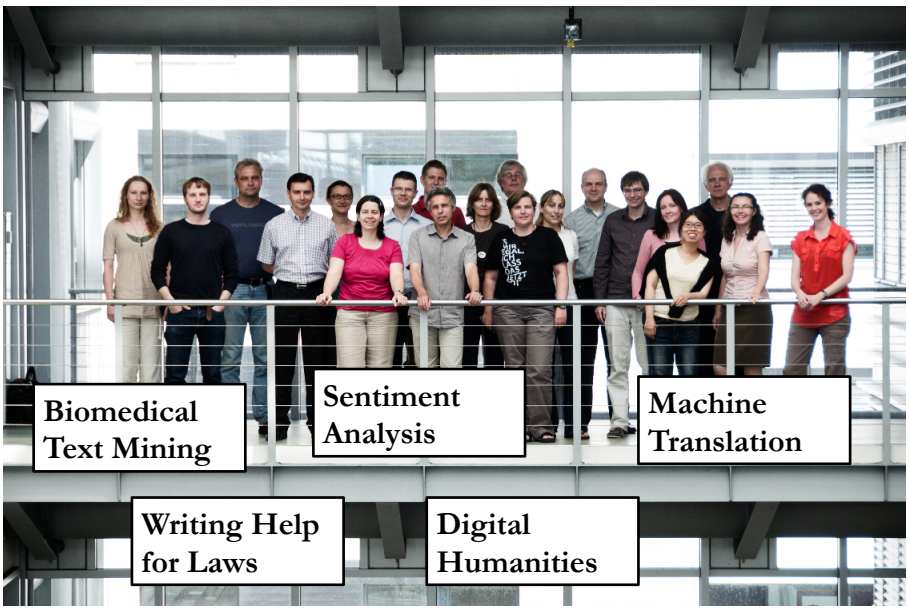
MACHINE TRANSLATION FOR SUBTITLES AND BEYOND

Martin Volk
Universität Zürich
volk@cl.uzh.ch



Subtitles – Audiovisual Translation

3

Subtitles

(by Diaz Cintas and Remael, 2007)

4

- Intralingual (same language as audio)
  - for the deaf and hard-of-hearing
  - for language learning purposes
  - for Karaoke effect
  - for dialects
  - for notices and announcements
- Interlingual (different language than audio)
  - for hearers
  - for the deaf and hard-of-hearing

20-Mar-14

Examples of Interlingual Subtitles

5

EN	DE
He's not picking up.	Er nimmt nicht ab.
So in order to save 100 bucks, you run the risk of permanent back injury.	Für $100 riskierst du also bleibende Rückenschäden.
- You're getting married? - When did you decide?	- Ihr werdet heiraten? - Wann habt ihr euch entschieden?
Look, I've been under a lot of stress lately…	Ja, ich hatte in letzter Zeit viel Stress
… and I've been releasing some tension, that's all.	und ich musste etwas von dem Druck loswerden, das ist alles.

20-Mar-14

Media Adaptation / Translation

6

- Subtitling
  - Netherlands, Portugal, Scandinavia, …
- Dubbing
  - France, Germany, Italy, Spain, …
- Voice Over
  - Eastern Europe, …
- Related to: Intertitling, Surtitling, Video Game Localization, Audio Description, …

20-Mar-14

Subtitling

- Different types of **production:**
  - Live or Real-time produced subtitles
    - Human made → respeaking
    - Automatic
  - Pre-produced subtitles
- Coverage
  - Full text / dialogue subtitles
  - Reduced subtitles

Automatic Translation of TV Subtitles

- a project by
  - our Computational Linguistics group together with
  - a large subtitling company
- 2006 – 2010
- language pairs:
  - Swedish → Danish
  - Swedish → Norwegian
  - English → Swedish

Translation Memory

Martin Volk

English	German
What's that all about?	Was soll denn das?
I can't wait for this day to be over.	Ich bin so froh, wenn dieser Tag vorbei ist.
Thank God it's only once a year.	Gott sei Dank ist er nur einmal im Jahr.
Maybe I'll meet him tomorrow.	Vielleicht werde ich ihn morgen treffen.
Chris wants to start up a band.	Chris möchte eine Band gründen.
…	…
…	…

Maybe I'll start up a band. → ???

Statistical Machine Translation

Martin Volk

English	German
What's that all about?	Was soll denn das?
I can't wait for this day to be over.	Ich bin so froh, wenn dieser Tag vorbei ist.
Thank God it's only once a year.	Gott sei Dank ist er nur einmal im Jahr.
Maybe I'll meet him tomorrow.	Vielleicht werde ich ihn morgen treffen.
Chris wants to start up a band.	Chris möchte eine Band gründen.
…	…
…	…

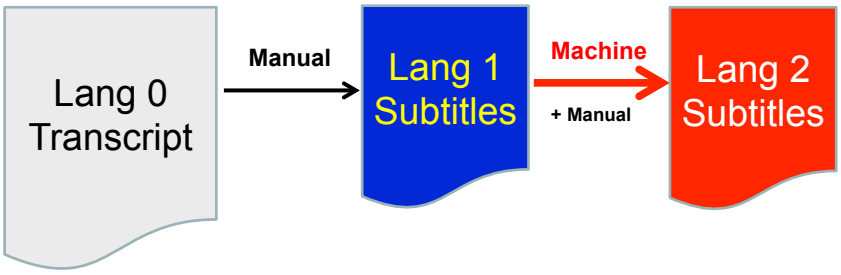
Statistical Machine Translation

Martin Volk

English	German
What's that all about?	Was soll denn das?
I can't wait | for this day to be over.	Ich bin so froh, | wenn dieser Tag vorbei ist.
Thank God | it's only | once a year.	Gott sei Dank | ist er nur | einmal im Jahr.
Maybe I'll | meet him tomorrow.	Vielleicht werde ich | ihn morgen treffen.
Chris wants to | start up a band.	Chris möchte | eine Band gründen.
…	…
…	…

Maybe I'll start up a band. → Vielleicht werde ich eine Band gründen.

Work Flow



Three Favorable Conditions for Automatic Translation

13

1. Subtitles are short.
   - Limited to two rows times 37 characters.
   - Average: ~ 10 words/subtitle
2. Closely related languages.
3. Huge amounts of human-translated high-quality subtitles.

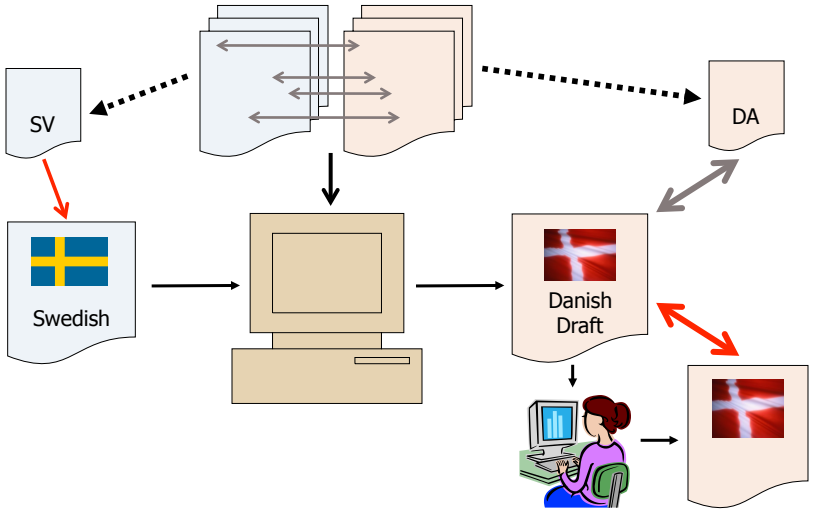
Unfavorable Conditions for Automatic Translation

14

- TV subtitles come from a large diversity of textual domains.
  - Comedy, Crime, Documentary, Talk show, …
- TV subtitles are a sort of (abbreviated) transcription of spoken language.
  - Emphatic spelling, Spacing, Bold-face
    - *lo-o-ove, Jerry S, e, i, n, f, e, l , d*
  - Many contractions
    - ***Nicole's*** *right.* ***We'd*** *appreciate your help.*

Evaluation Scenario

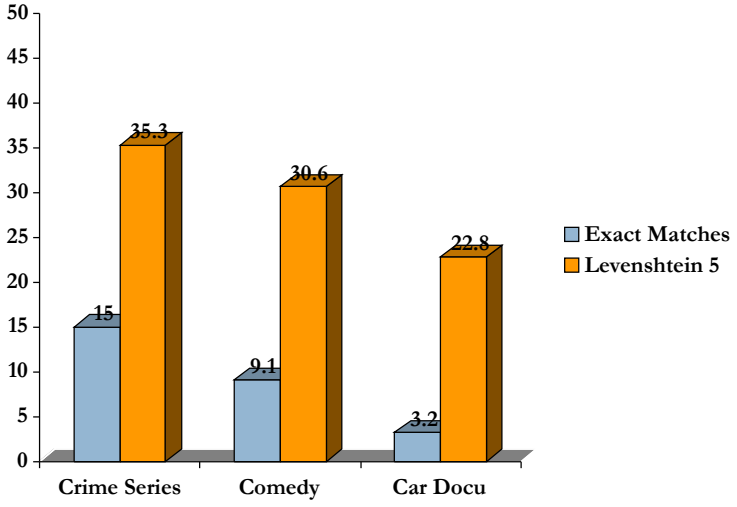
15



Evaluation Results against a Prior Human Translation

(for Swedish →Danish)

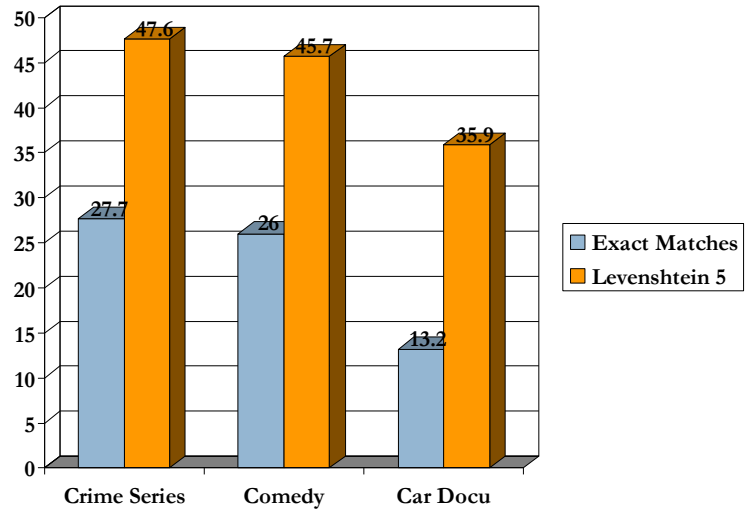
16



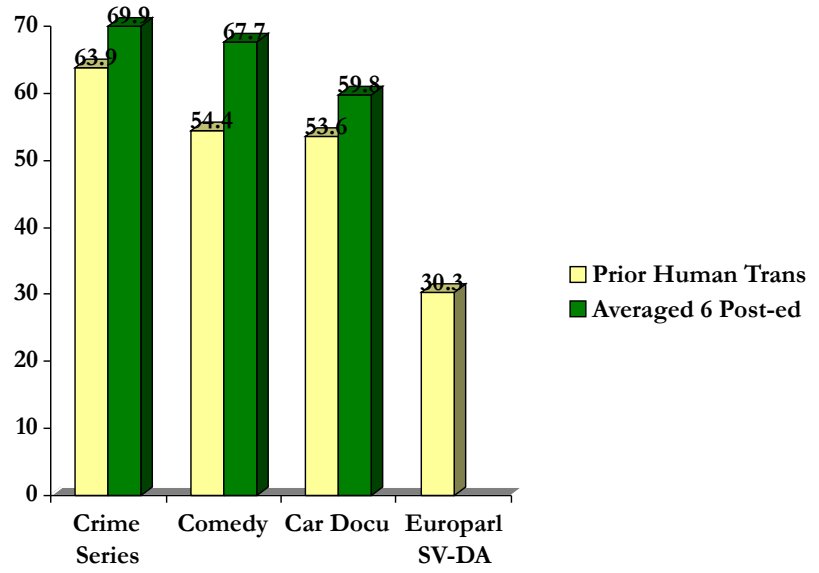
Evaluation Results averaged over 6 Post-editors

(for Swedish →Danish)

17



Evaluation Results in BLEU scores



20-Mar-14

18

19

Source Language	Target Language										
	da	de	el	en	es	fr	fi	it	nl	pt	sv
da	-	18.4	21.1	28.5	26.4	28.7	14.2	22.2	21.4	24.3	28.3
de	22.3	-	20.7	25.3	25.4	27.7	11.8	21.3	23.4	23.2	20.5
el	22.7	17.4	-	27.2	**31.2**	**32.1**	11.4	26.8	20.0	27.6	21.2
en	25.2	17.6	23.2	-	**30.1**	**31.1**	13.0	25.3	21.0	27.1	24.8
es	24.1	18.2	28.3	**30.5**	-	**40.2**	12.5	**32.3**	21.4	**35.9**	23.9
fr	23.7	18.5	26.1	**30.0**	**38.4**	-	12.6	**32.4**	21.1	**35.3**	22.6
fi	20.0	14.5	18.2	21.8	21.1	22.4	-	18.3	17.0	19.1	18.8
it	21.4	16.9	24.8	27.8	**34.0**	**36.0**	11.0	-	20.0	**31.2**	20.2
nl	20.5	18.3	17.4	23.0	22.9	24.6	10.3	20.0	-	20.7	19.0
pt	23.2	18.2	26.4	**30.1**	**37.9**	**39.0**	11.9	**32.0**	20.2	-	21.9
sv	**30.3**	18.9	22.8	**30.2**	28.6	29.7	15.3	23.9	21.9	25.9	-

Table 2: BLEU scores for the 110 translation systems trained on the Europarl corpus

taken from: Philipp Koehn: *Europarl: A Parallel Corpus for Statistical Machine Translation*. MT-Summit. 2005.

20-Mar-14

Interview with Post-editors

20

- interviews with two post-editors
- generally positive impression
- post-editing is different from translating
- post-editing is hard cognitive work
  - danger of being lead in a particular direction
- estimated time-saving: 20-30%
- big differences between post-editors (generous vs. picky).

Martin Volk, Uni Zurich 1. April 2011

Time Saving Study

21

- by Sheila C. M. de Sousa,Wilker Aziz and Lucia Specia (University of Wolverhampton, 2011)
- English → Portuguese
- Google Translate vs. Own MT vs. Translation from Scratch
- Large experiment with 11 "volunteers"
  - native speakers of Brazilian Portuguese
  - "some experience with translation tasks"

→ Postediting is 40% faster than Translation from Scratch!

Subtitle Translation vs. Sentence Translation

22

Master Project by Susanne Rozkosny at ZHAW
- with supervision by Gary Massey

- Subtitle-based MT (EN → DE)
  - Look, I've been under a lot of stress lately…
  - … and I've been releasing some tension, that's all.
- Sentence-based MT (EN → DE)
  - Look, I've been under a lot of stress lately and I've been releasing some tension, that's all.

→ Better quality for subtitle-based MT!

EU-Project:

www.sumat-project.eu

- coordinated by VicomTech, San Sebastian, Spain
- with technical partners in Greece, Ireland, Slovenia, Switzerland
- with subtitling companies
  - Deluxe Digital Studios
  - Invision Ondertiteling
  - Titelbild Subtitling and Translation
  - Voice & Script Int. (VSI)
- 2011-2014
- EN - DE, ES, FR, NL, PT, SV plus Slovenian and Serbian

English – Swedish MT in SUMAT

Parallel Subtitles as Training data

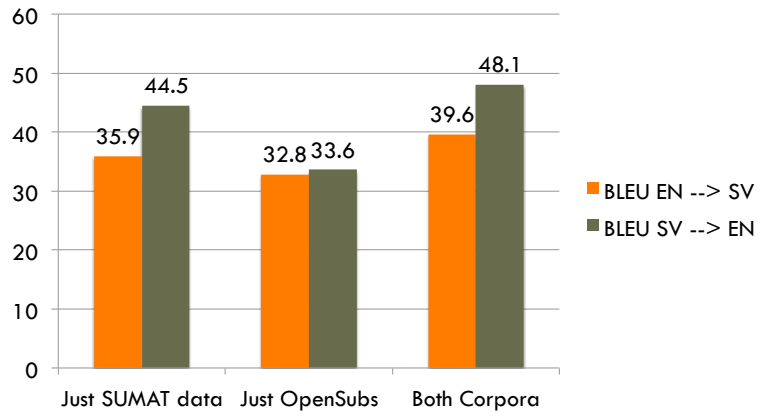
- SUMAT data (contributed by the subtitle companies)
  - 548,000 EN – SV subtitle pairs
  - 5.6 million EN tokens / 4.6 million SV tokens
- Open Subtitles (freely available on the web)
  - 7.2 million EN – SV subtitle pairs
  - 55.3 million EN tokens / 48.1 million SV tokens

Monolingual data for the Language Model

- English: 8.1 million subtitles / 76.3 million tokens
- Swedish: 3.7 million subtitles / 37.8 million tokens

English – Swedish MT in SUMAT

25



20-Mar-14

Quality Estimation

26

- = Confidence Estimation
- = The SMT system delivers only translations when confident, otherwise nothing.
- Quality Estimation is trained as a classifier based on features such as
  - subtitle length
  - number of unknown words
  - number of translation options
  - …

20-Mar-14

Conclusions

- Subtitling is an interesting field for translators.
- Subtitles are well-suited for Statistical Machine Translation.
- Statistical Machine Translation enables you to re-use your translation archive and to profit from previous work.

Domain-specific SMT

Text+Berg digital
Projekt zur korpuslinguistischen Erschliessung alpinistischer Literatur
www.textberg.ch

Goal: SMT systems for Alpine texts (DE-FR)

Based on the Text+Berg corpus

- SAC texts from 1864 – 2011
- Parallel part 1957 – 2011: ~6 million tokens

Domain-specific SMT

- Source: 20. Juni : unser dritter Angriff auf das Gross Grünhorn (4044 m) .
  Human: Le 20 juin eut lieu notre troisième tentative au Gross Grünhorn (4044 m) .
- Google: 20 . Juin : notre troisième attaque majeure sur la Corne de Green (4044 m) .
- We: 20 juin . notre troisième tentative vers le Gross Grünhorn (4044 m) .

Domain-specific SMT: Research Questions

1. How can we combine **large out-of-domain corpora** with small domain-specific corpora?
2. How can we combine **large monolingual corpora** on the source or target side with small domain-specific corpora?
3. How can we use domain-specific **terminology** to improve an SMT system?

Domain-specific SMT: Results

31

	German → French	BLEU	French → German	BLEU
1	Google Translate:	12.95	Google Translate:	10.53
2	Personal Translator 14:	13.29	Personal Translator 14:	08.93
3	Moses (Europarl data):	11.46	Moses (Europarl data):	09.27
4	Moses (SAC-145 data):	16.91	Moses (SAC-145 data):	14.72
5	+ Bleualign:	18.18	+ Bleualign:	15.47
6	+ Reordering:	18.38		
7	+ Lattice:	18.80		
8	+ Interpolated LM	18.98	+ Interpolated LM	15.69
9	+ Combined Translation Model	19.13	+ Combined Translation Model	15.81
10	+ Dynamic Translation Model	20.24	+ Dynamic Translation Model	16.69

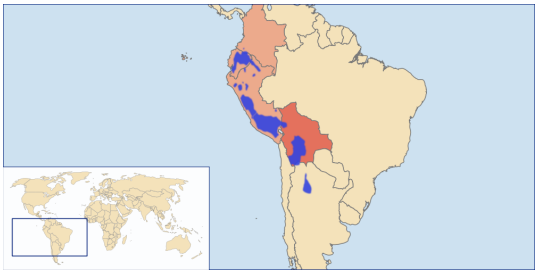
Hybrid Machine Translation for Lesser-resourced languages

32

- We are investigating

Spanish → Quechua Machine Translation

- Funded by SNF 2011-2015



Bilingwis ES → QU

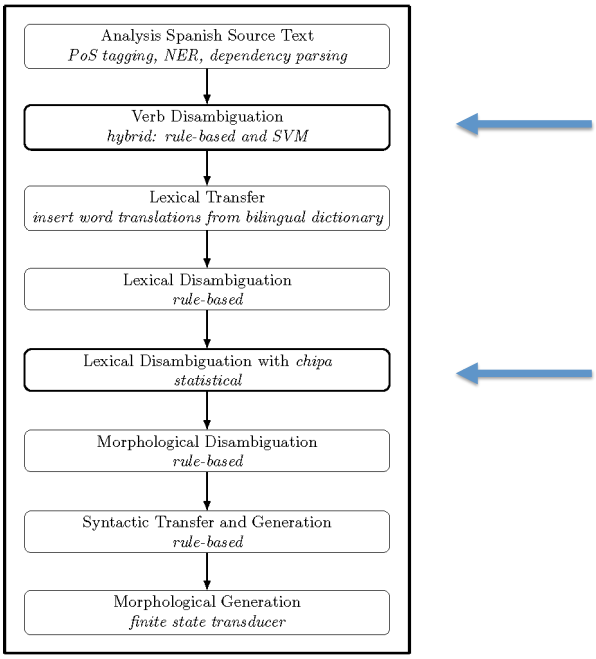
33

sunqu — 21 hits

Search for 'sunqu'

Entonces, desde ese día, en mi **corazón** se prendió, como alfiler, la idea de salir de la casa de mi madrina para ir abuscar trabajo.	Chayqa chay p'unchaymanta pacha madrinaypa wasinmanta ripunayachikuyniy **sunqu**ypi aguja hina t'iparayan trabajo maskhakuq rinaypaq.
Ojalá aesta señora de buen **corazón** el Señor la haya hecho sentar asu lado, porque ella es la que me salvó de lo que ya estaba caminando ala otra vida.	Ojala chay sumaq **sunqu**yuq señorata Taytanchis ladonpi tiyachishanman, paymi salvawaran ña huq kaq vidaman puririshaqtiyña.
En lo que estaba caminando así tras las ovejas, con el **corazón** puesto ya en otro pueblo, un día pasaron unos arrieros con dirección aAcopia.	Chhayna uwihakunaq qhipan purishasqaypi, ña huq lado llaqtapi **sunqu**yuqña, huk p'unchay arrierokuna pasanku Aqopiya ladoman.
Así, en lo que estaba caminando entre las q'oyas, llorando ypenando mi suerte amarga como la sal, una mujer de buen **corazón** me llevó: Ya no llores, hace mucho rato se han ido, me dijo.	Chayqa quya ukhupi kachi vidaymanta llakispa waqakuspa purishaqtiy, huk allin **sunqu**yuq warmi pusawan.

20-Mar-14



20-Mar-14

34

MT for the Automobile domain (KTI project 2012-2013)

35

- Partner: Translation Company Semiotic Transfer (Baden, AG)
- Languages: DE → FR and DE → IT
- Texts: Marketing Material for Automobile Industry

semiotic transfer

La nouvelle Golf Plus Highline.

Élégance
Elle confère une touche d'élégance à la grisaille du quotidien: la nouvelle Golf Plus Highline. On le voit aux détails chromés tels que la calandre noire avec baguettes chromées, les phares antibrouillard avec encadrement chromé et éclairage en virage ou encore l'encadrement chromé de la grille de prise d'air. Les jantes en alliage léger 16" au design "Cleveland" donnent un accent particulier au véhicule. Enfin, les pare-chocs, boîtiers de rétroviseurs extérieurs et poignées de portes sont naturellement laqués de la couleur du véhicule afin de parfaire l'impression d'ensemble.

zudem die 16-Zoll-Leichtmetallräder im Design „Cleveland". Für einen stimmigen Gesamteindruck sind Stoßfänger, Außenspiegelgehäuse und Türgriffe selbstverständlich in Wagenfarbe gehalten.

MT for Sign Language

36

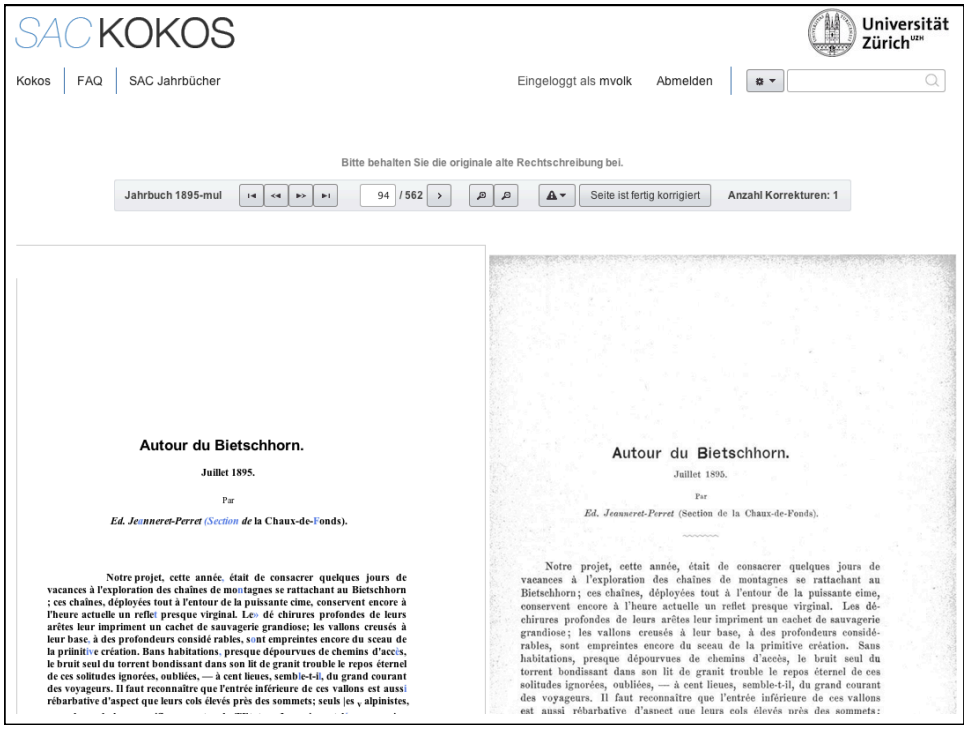
- Translation: German → Swiss German Sign Language
- via Sign Language Notation + Avatar
- Domain: Railway Announcements

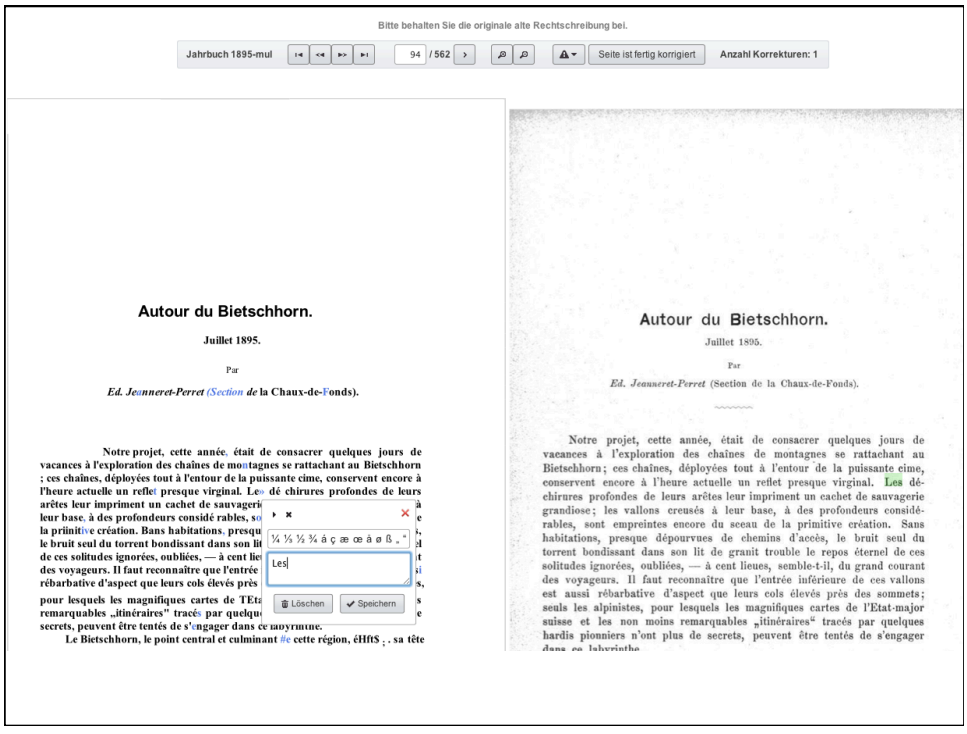
Example: "*Dieser Zug endet in Brig*"

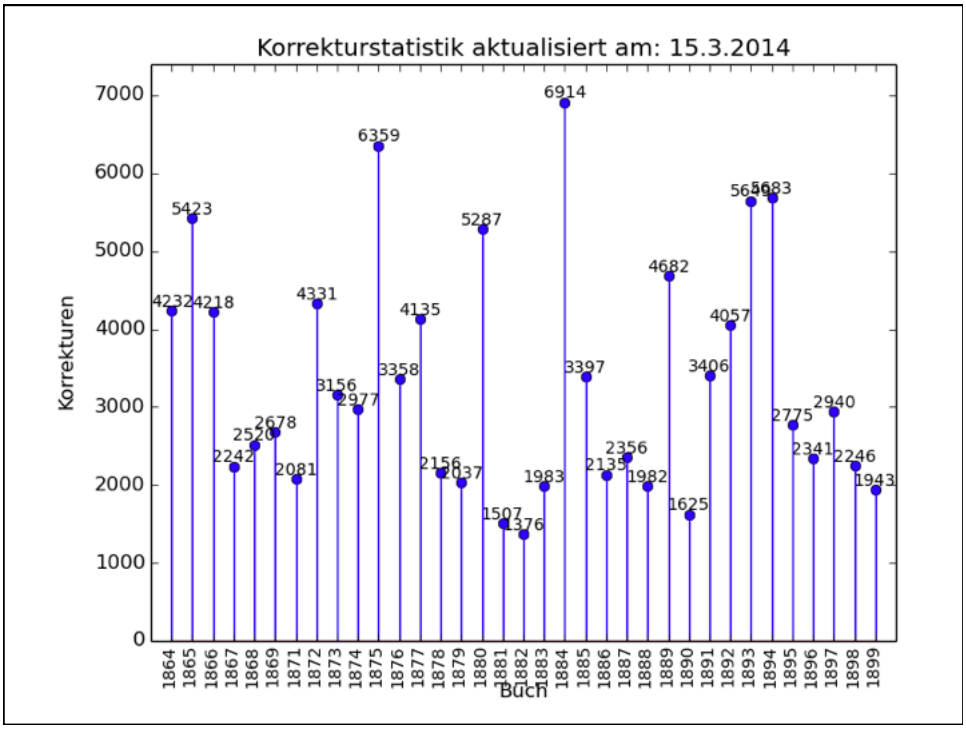
→→→

→→→

20.03.14


SAC KOKOS
Universität Zürich
Kokos
FAQ
SAC Jahrbücher
Eingeloggt als mvolk
Abmelden
Bitte behalten Sie die originale alte Rechtschreibung bei.
Jahrbuch 1895-mul
94 / 562
Seite ist fertig korrigiert
Anzahl Korrekturen: 1
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 reflet 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,


Bitte behalten Sie die originale alte Rechtschreibung bei.
Jahrbuch 1895-mul
94 / 562
Seite ist fertig korrigiert
Anzahl Korrekturen: 1
Autour du Bietschhorn.
Juillet 1895.
Par
Ed. Jeanneret-Perret (Section de la Chaux-de-Fonds).
Les
Löschen
Speichern
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 reflet 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'État-major suisse et les non moins remarquables „itinéraires" tracés par quelques hardis pionniers n'ont plus de secrets, peuvent être tentés de s'engager



Please join us!

Correct some errors

in the old yearbooks of the SAC

in Kokos

[**Ko**llaboratives **Ko**rrektur**s**ystem]

http://kokos.cl.uzh.ch