Credit Suisse Bulletin Corpus: The world's oldest banking magazine as a treasure trove of applications for Digital Humanities

Gerold Schneider, Anastassia Shaitarova and Martin Volk
{gerold.schneider,anastassia.shaitarova,martin.volk}@uzh.ch
University of Zurich[UZH]

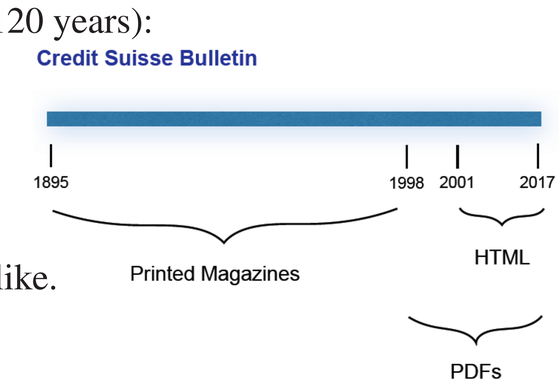
DARIAH Workshop 2018, Neuchâtel

Introduction

We present a diachronic parallel corpus on the Credit Suisse Bulletin, the world's oldest banking magazine, and three of its applications: vocabulary changes, the translation of the partitive particle, and trends and changes in society over time.

CS Bulletin Corpus

THE CREDIT SUISSE BULLETIN CORPUS has been published since 1895 in German, with translations in French and partly in English and Italian. The German and French part contain over 20 million words each, the English and Italian part about 10 million words each. Our parallel, multilingual corpus fills a gap in parallel corpora with respect to genre (magazine articles), domain (banking and economy articles), and its time span (120 years):



It can be freely accessed and downloaded. Its characteristics allow researchers to conduct a multitude of orthogonal types of research, turning it into a treasure for historians, sociologists, and linguists alike. We showcase three applications:

Application: Vocabulary

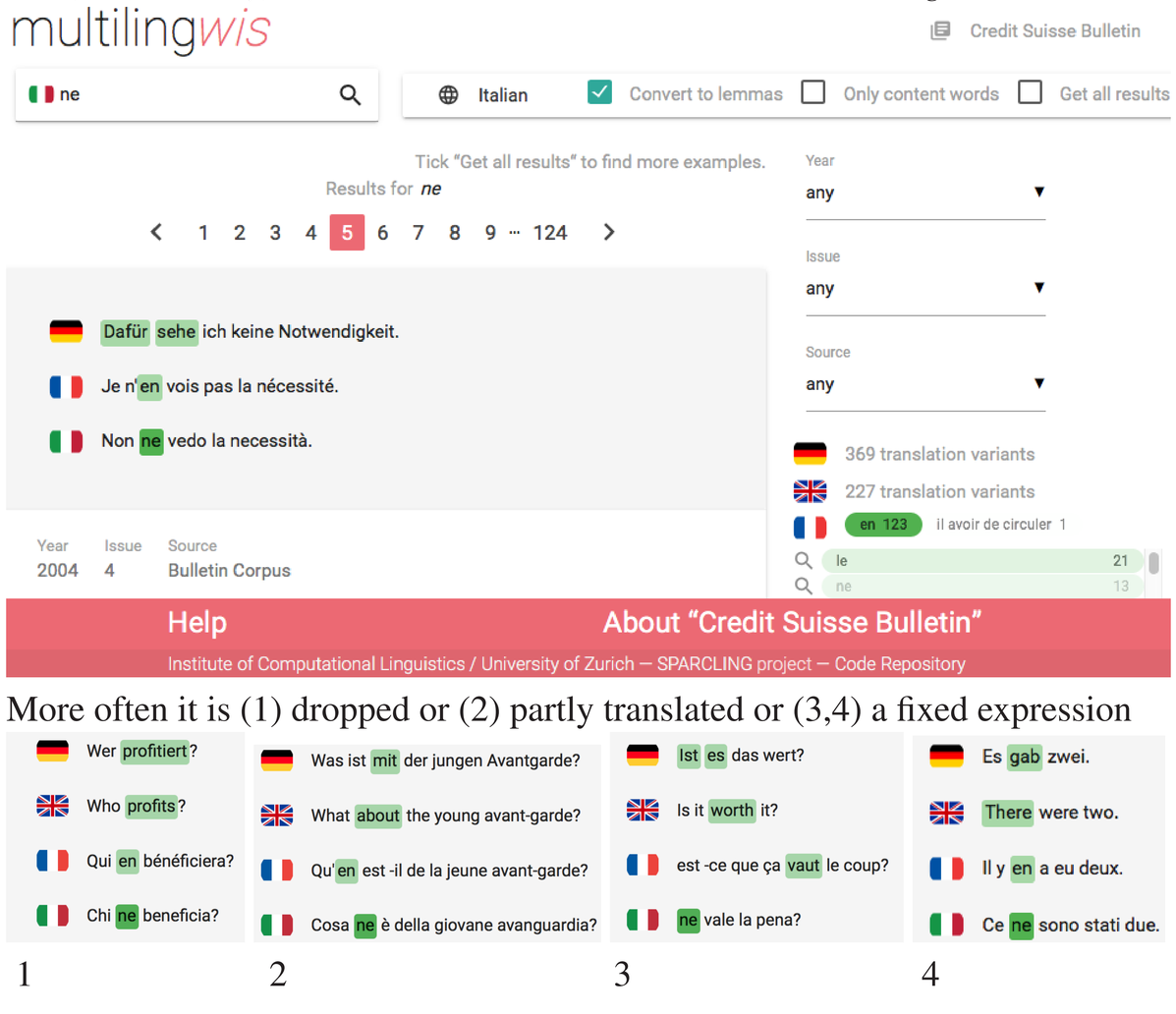
DOCUMENT CLASSIFICATION is applied to the corpus split into an early (until 1969) and late (from 1970 on) period. We use logistic regression on a monogram bag-of-word model. The classification accuracy is 98.5%. Many of the features show developments in banking vocabulary, language, orientation of the magazine. We show selected features ranked by decreasing weight.

Late Period				Early Period			
Rank	word	Freq	weight ↓	Rank	word	Freq	weight ↓
4	zeitfrage	1129	17.001	5	krieg	700	6.893
5	foto	2729	16.945	6	aktienindex	163	6.674
8	bulletin	6445	14.363	7	finanziell	1165	6.527
11	anleger	2703	13.893	9	reingewinn	1691	6.482
16	kunde	2464	11.761	13	indessen	1053	6.234
17	suisse	4066	11.192	14	aktiengesellschaft	1772	6.069
20	partner	1053	10.737	16	bern	1453	5.951
26	ordern	28	10.347	19	wechselstube	349	5.644
28	weltweit	3045	10.184	20	industrie	2141	5.615
30	perspektive	1651	9.837	22	zinsfuss	276	5.553
32	spreitenbach	186	9.769	25	herr	188	5.403
37	warenhaus	177	9.086	31	bundesbahn	902	5.198
42	mio	1069	8.569	38	million	3422	5.085
43	zinstrend	181	8.505	39	valuta	470	5.053
44	grafik	822	8.473	40	westdeutschland	219	4.976
45	veränderung	1975	8.471	41	bankier	60	4.917
46	welt	3773	8.389	43	pfandbrief	319	4.867

Selected features (see Rank) of early and late period.

Application: Translation

THE ITALIAN PARTITIVE PARTICLE *ne* is translated to French *en* in about half of the instances, and direct translations to German are rarer (*dafür, davon, dazu, da, daraus*), as one can see in our online browser *multilingwis*:



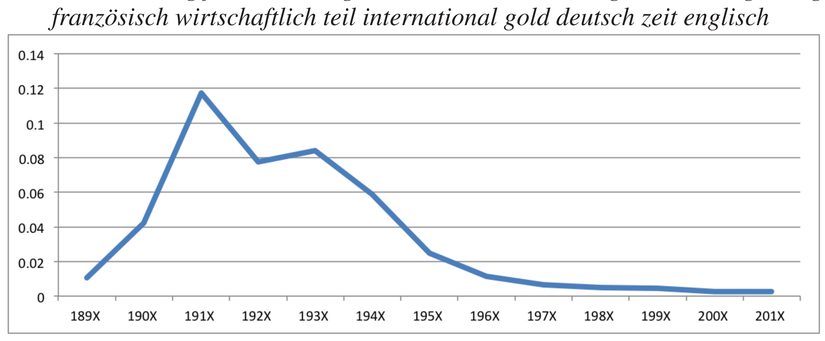
More often it is (1) dropped or (2) partly translated or (3,4) a fixed expression

Application: Changes in Society

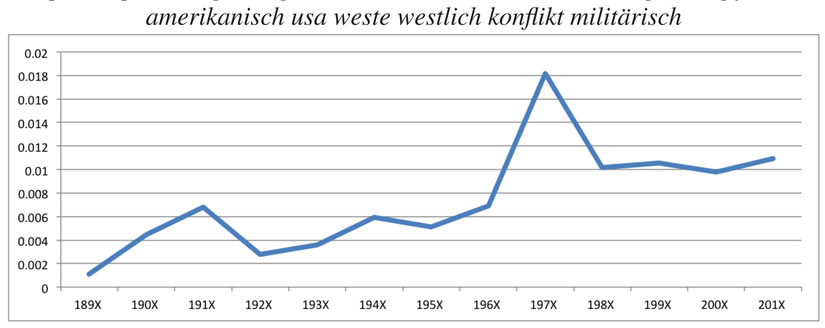
TOPICS IN SOCIETY are changing, and also what a modern banking magazine wants to be seen as reporting on.

We use topic modeling to detect hidden topics and investigate how they change over time. With 50 topics we observe e.g. the following changes in selected topics (we give keywords and proportion of decade):

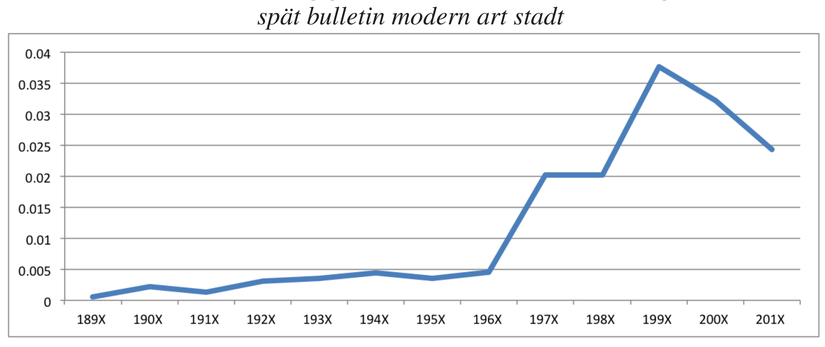
land staat milliarde million krieg frankreich england deutschland vereinigt ausland regierung amerikanisch französisch wirtschaftlich teil international gold deutsch zeit englisch



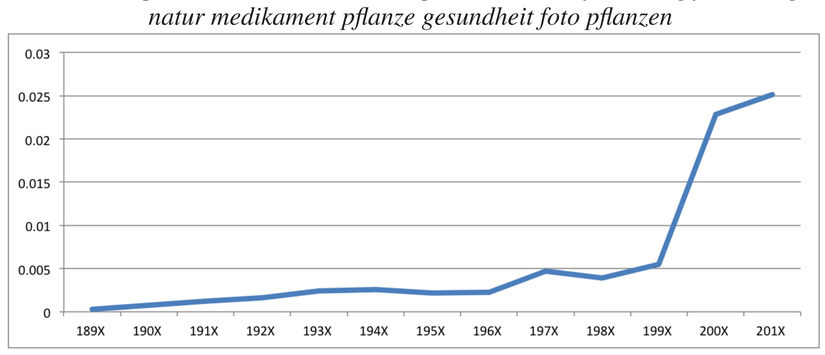
politisch präsident politik partei regierung wahl staat land osten westen europa krieg frankreich sowjetunion amerikanisch usa weste westlich konflikt militärisch



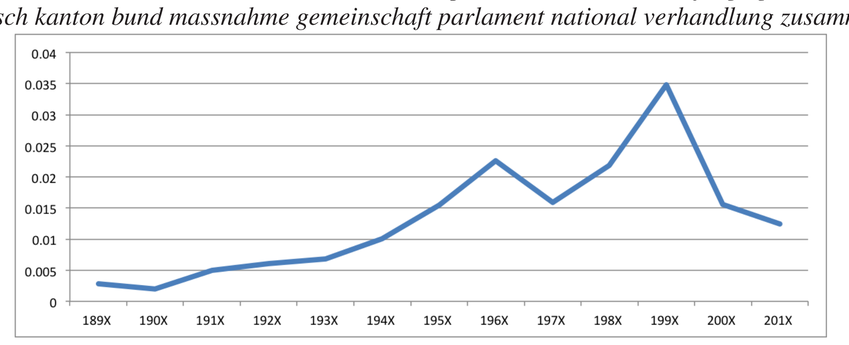
jahrhundert kunst werk bild künstler ausstellung geschichte alt museum sammlung zeit münze magazin zürich buch spät bulletin modern art stadt



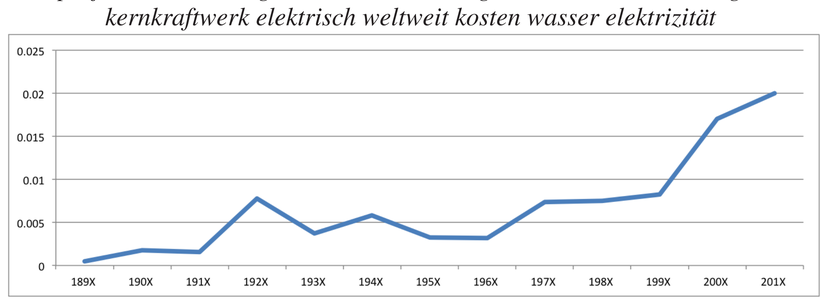
mensch tier wein arzt wasser patient essen krankheit körper menschlich forschung forscher gesund medizinisch natur medikament pflanze gesundheit foto pflanzen



europäisch schweiz land staat bundesrat international politisch schweizerisch frage gemeinsam wichtig eidgenössisch kanton bund massnahme gemeinschaft parlament national verhandlung zusammenarbeit



energie bau bauen projekt wald strom gebäude umwelt anlage erneuerbar nachhaltig stadt holz ökologisch kernkraftwerk elektrisch weltweit kosten wasser elektrizität



Conclusions

- The CS Bulletin corpus is a multi-purpose corpus.
- As a diachronic multilingual parallel corpus it allows statistical analyses of changes in language and society over the span of 120 years.
- Aligned sentences serve as data set for translation tools.
- We have shown how text classification and topic modeling tools can be used on our corpus to visualize societal trends.

References

Our corpus can be downloaded at http://pub.cl.uzh.ch/projects/b4c/en/corpora.php
Volk/Amrhein/Aepli/Müller/Ströbel. 2016. "Building a Parallel Corpus on the World's Oldest Banking Magazine". *Proceedings of KONVENS 2016*, Bochum.