SPARCLING: Large-scale Annotation and Alignment of Parallel Corpora for the Investigation of Linguistic Variation

Marianne Hundt; Martin Volk; Elena Callegaro; Johannes Graën
University of Zurich, English Department & Institute of Computational Linguistics

ABSTRACT

This collaboration between the departments of English and Computational Linguistics builds on the use of parallel multilingual corpora.

The project aims at annotating (PoS-Tagging and Parsing) and aligning (both sentence and word alignment) a large parallel corpus, the *Europarl Parallel Corpus*, for the following languages pairs: English-German, English-French and English-Spanish. Additional language pairs to be added later (from the MultiUN corpus): English-Russian and English-Finnish.

The project sets out to prove the usefulness of such large annotated and word-aligned corpora for the investigation of linguistic variation. In particular, contexts where one language uses a structure (e.g. an article) are used to retrieve 'zero' contexts in other language(s).

The computational linguistic task is to build an efficient, powerful and highly innovative corpus query tool that is able to handle these corpora.

CONTACTS

Prof. Dr. Marianne Hundt
English Department
Email: m.hundt@es.uzh.ch

Prof. Dr. Martin Volk
Institute of Computational Linguistics
Email: volk@cl.uzh.ch

Europarl Transcriptions as 'Corpus'

A case study on recordings vs. transcriptions
The following are the results of the comparison between the videos and the transcriptions:
a. Some elements are removed:
- greetings and thanks;
- obvious mistakes;
- repetitions;
- linguistic forms that are perceived as 'spoken' (e.g. contractions);

b. There is a tendency to use a more standard and formal language, which is closer to the written language.

Type	Transcription	Video
Word order	Vorab möchte ich allen meine Gratulation zur erreichten Einigung **aussprechen**!	Vorab möchte ich allen meine Gratulation **aussprechen** zur erreichten Einigung!
	Unfortunately, what we see and what we have **here** on the table…	Unfortunately, what we see and what we have on the table **here**…
Deletion or modification of words	I recall clearly, as **the rapporteur** dealing with the EU-India Free Trade Agreement…	I recall clearly, as **rapporteur** dealing with the EU-India Free Trade Agreement…
	When one looks at the situation with regard to **the growth** in applications, it is the same.	When one looks **at growth** in applications, the situation there once again.
Deletion or modification of informal expression and/or words	You **cannot**.	You **can't.**
	So why does it matter? As has been mentioned, we have for several decades had the European Patent Office **which** is an international organisation outside the EU.	So why does it matter? **Well**, as has been mentioned, we have for several decades had the European Patent Office **that** is an international organisation outside the EU.
Deletion of foreign words	Forschung	Research
	Übereinkommen	Agreement

Table 1. Comparison between speech recordings and transcriptions of the European Parliament.

Variable article use

Hypothesis: Articles are used more frequently in German than in English, and using a parallel corpus of German and English is therefore a good way of targeting bare NPs in English.
Two case studies:
Parlament/parliament (non-aligned data):

***Parlament/parliament* in Europarl**	**English (count)**	**German (count)**
Definite article	86 (43%)	166 (83%)
Bare NP	99 (49.5%)	4 (2%)
Other determiner (demonstrative or possessive)	15 (7.5%)	30 (15%)

Sicherheit/safety (sentence-aligned data):

***Sicherheit/Safety* in Europarl**	**Count**
Null article in English and German	7 (14%)
Definite Article in English and German	10 (20%)
Null article in English only	32 (64%)
Null article in German only	1 (2%)

Chart 1. Comparison in variable article use of the words *Parlament/parliament* in English and in German.

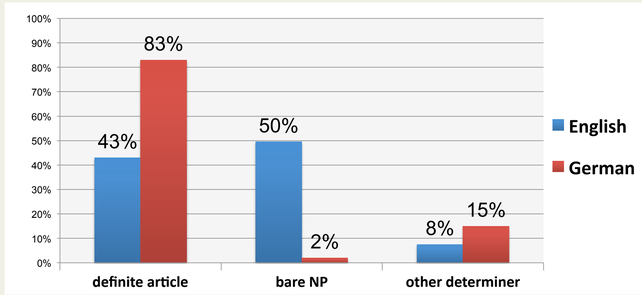
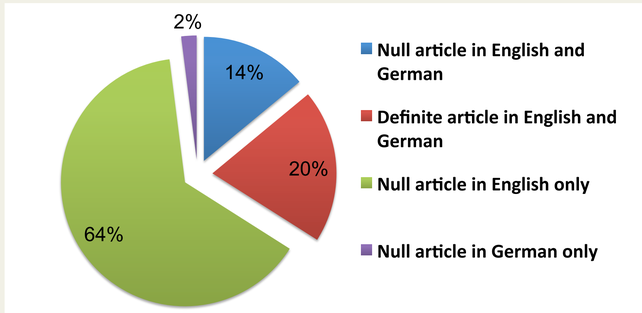
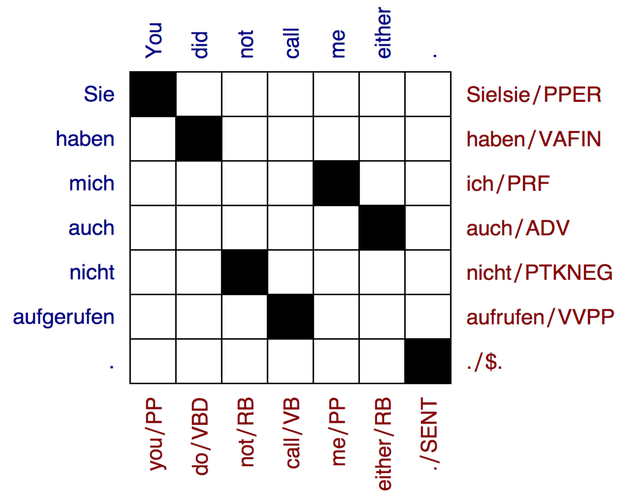


Chart 2. Comparison in variable article use of the words *Sicherheit/safety* in English and German.



Annotation and alignment

Potential sentence boundaries have been identified by combining PoS tags and meta information from the compiled corpus data (Koehn, 2005). By applying hunalign (Varga, 2005) to every language pair out of five languages that we could process with the TreeTagger (Schmid, 1994) and by using pre-trained tagger parameter files, we were able to "harmonize" the resulting alignments by removing those which were not covered by other alignments transitively and hence probably erroneously.
In the next step, we will provide word alignment for all aligned sentences and expand our query tool to these.



REFERENCES

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT summit*, vol. 5, pages 79-86.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, vol. 12, pages 44-49, Manchester, UK.

Varga, Dániel and Nemeth, László and Halácsy, Péter and Kornai, András and Trón, Viktor and Nagy, Viktor. 2005. Parallel Corpora for Medium Density Languages. In *Proceedings of the RANLP*, pages 590-596.

Rafalovitch, Alexandre and Robert, Dale. 2009. United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. *Proceedings of the MT Summit*, vol. 12, pages 292-299.

http://www.europarl.europa.eu/sides/getDoc.dotype=PV&reference=20121211&secondRef=TOC&language=en

http://www.cl.uzh.ch/research/parallelcorpora/sparcling.html

http://www.es.uzh.ch/Subsites/Projects/SPARCLING.html