What’s a Collocation?

- "recurrent combinations of words that co-occur more often than expected by chance and that correspond to arbitrary word usages."
- My definition: Words that are used together a lot. They often have a meaning that is different than the sum of their parts.

What's the project about?

- Reimplementations of two previous papers
- Xtract, a collocation extractor
  - Uses statistical methods to find collocations
- Champollion, a collocation translator
  - Uses statistical methods to translate collocations

Implementation Details

- Written in Java
- Uses Lucene
  - Indexing for speed optimization
- Uses the Europarl Corpus
  - English-German sentence-aligned proceedings of the European Parliament

JXtract

- Accepts a word and a corpus as input, and it outputs collocations
- Found the most frequently used words in the corpus
- Gave most frequent words to JXtract to find collocations

JXtract Sample Output

```
the european union _ most
most of us
the most vulnerable
the report presented
report on competition policy
the most _ _ in the world
a number of _ have already
the committee on citizens ' freedoms and rights justice
european citizens
our fellow citizens
the committee on agriculture and rural development
the committee on constitutional affairs
the committee on budgetary control
the committee on development and cooperation
the committee on culture youth education the
as far as possible
certain member states
```

JChampollion

- Accepts a sentence-aligned, bilingual corpus and a collocation in the source text
- Produces a translation of the collocation in the target language
- Originally English-French, I used English-German
- Europarl Corpus
- Apache Lucene indexer

JChampollion Sample Output

Source Text	Jchampollion Translation
Madam President	frau präsidentin
member states	mitgliedstaaten
the committee on agriculture and rural development	landwirtschaft ländliche
report on competition policy	wettbewerbspolitik

Limitations

- JXtract does not use tag and parse information like Xtract does
- Champollion never outputs closed class words (prepositions, articles) as part of the translations
  - They are so frequent they mess up the statistical correlations, so they must be excluded
- It’s slow
  - Both programs the order of one minute per collocation

References

- Smadja, F. 1993. Retrieving collocations from text: Xtract. *Comput. Linguist.* 19, 1 (Mar. 1993), 143-177.
- Smadja, F., McKeown, K. R., and Hatzivassiloglou, V. 1996. Translating collocations for bilingual lexicons: a statistical approach. *Comput. Linguist.* 22, 1 (Mar. 1996), 1-38.
- Europarl: A Multilingual Corpus for Evaluation of Machine Translation, Philipp Koehn, Draft, Unpublished
- http://lucene.apache.org/