Semantic Shifts

Semantic Shifts: From lexicon to grammar. Diachronic and typological perspectives

Porquerolles
14-25 Sept. 2020

Tracing language change using parallel corpora

Robert Östling

Parallel text is an underutilized resource for large-scale studies in linguistics. This may be due to the fact that manual analysis at this scale is too time-consuming, while the technical presentations of automated methods in the computational linguistics literature may feel intimidating.

However, digital parallel corpora have become too useful to ignore. These come in many different forms, suitable for a range of purposes. Long texts, typically government-produced, are well-suited for more detailed studies of the few languages they are available in. At the other end of the spectrum, shorter texts, often religious, are available for many more languages and allow world-wide typological studies. Some corpora contain multiple translations into the same language from different time periods, and can be applied directly to diachronic studies. More commonly however, parallel texts provide a rich set of synchronous data that can be used by traditional methods for inferring historical developments.

This course contains two sessions of 75 minutes.

Session 1: tools and resources
In the first session, we will cover important parallel corpora and tools used for linguistic research. First we will have a closer look at a few important corpora that are frequently used in linguistics research: the Europarl corpus (Koehn 2005), the Parallel Bible and Watchtower corpora (among others, see Mayer and Cysouw 2014). Then, we proceed with some basic models and tools used for analyzing parallel text. The IBM models and derivatives have become a de-facto standard for word level alignment (Och and Ney 2003), and many tools are available (e.g. Östling and Tiedemann 2016). For many purposes, mathematically simpler co-occurrence based methods can provide useful results.

Session 2: possibilities and problems
In the second session, we will look at how the corpora and tools discussed in the first session can be applied. Examples include lexical typology (Östling 2016) and the typology of word order in adjectives (Östling and Wälchli 2019), numerals (Kann 2019) and more generally (Östling 2015). These examples, and many others, hint at the possibilities. However, while some of the parallel text methods may seem like magic at first, they are not. To avoid disappointment one must be aware of the limitations inherent in different methods. These come at all levels, from mathematical simplifications in word alignment algorithms to the translation history of the Bible.

Literature and references:

Two of the references below are particularly useful as starting points. The special issue introduced by Cysouw and Wälchli (2009) is still a relevant summary of using parallel corpora in linguistic typology.

Ponti et al. (2019) provide a more up-to-date summary, although from a different point of view as their focus is on practical applications in natural language processing.

The remaining references concern specific resources, methods or studies and are included here for interested researchers to choose from. There is a distinct bias towards my own work, because there I am most familiar with the technical details that frequently turn out to be essential in this line of work.

Östling, Robert and Wälchli, Bernhard (2019). Word-order goes lexical typology: Adjective-noun order and massively parallel text (abstract). 13th Conference of the Association for Linguistic Typology. September 2019.

Östling, Robert. The Lexical Typology of Semantic Shifts, chapter Studying colexification through massively parallell corpora, pages 157--176. De Gruyter, 2016.
http://dx.doi.org/10.1515/9783110377675-006

Östling, Robert and Tiedemann, Jörg. Efficient word alignment with Markov Chain Monte Carlo. Prague Bulletin of Mathematical Linguistics, 106:125--146, October 2016. http://ufal.mff.cuni.cz/pbml/106/art-ostling-tiedemann.pdf

Och, Franz Josef and Ney, Hermann (2003). A systematic comparison of various statistical alignment models. Computational Linguistics. 29 (29): 19-51.

Cysouw, Michael and Wälchli, Bernhard (2009). Parallel texts: using translational equivalents in linguistic typology. STUF - Sprachtypologie und Universalienforschung, 60(2), pp. 95-99.

Koehn, Philipp (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit 2005.

Mayer, Thomas & Cysouw, Michael (2014). Creating a Massively Parallel Bible Corpus. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). May 2014.

Östling, Robert (2020). parallel-tools: a software package for simple searches in massively parallel texts.
https://github.com/robertostling/parallel-tools

Östling, Robert (2015). Word order typology through multilingual word alignment. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 205--211, Beijing,China, July 2015. Association for Computational Linguistics.
http://www.aclweb.org/anthology/P15-2034

Ponti, Edoardo Maria and Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, Anna Korhonen (2019). Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. Computational Linguistics 2019 45:3, 559-601
https://doi.org/10.1162/coli_a_00357