A cookbook of co-occurrence comparison techniques and how they relate to the subtleties in your research question

Viola Wiegand (University of Birmingham, UK), Anthony Hennessey (University of Nottingham, UK), Christopher R. Tench (University of Nottingham, UK) and Michaela Mahlberg (University of Birmingham, UK)

Introduction

The concept of 'collocation' is one of the most fundamental in corpus linguistics. Its centrality is reflected in the range of definitions and approaches that have been developed in corpus linguistics over the past decades (for detailed accounts see e.g., Evert, 2008; Gries, 2013; McEnery & Hardie, 2012). The observation and quantification of collocations has been crucially used for disambiguating different senses of words, for example, in the COBUILD dictionary project (Sinclair, 1987). Beyond the application in lexicography, the description of collocations is also relevant, among others, in the contexts of learner corpus research and discourse analysis. Although corpus linguistic approaches are inherently comparative, only a few studies have taken the analysis and interpretation of collocation beyond a single corpus. For example, Gabrielatos and Baker (2008: 11) identify 'consistent collocates' across annual newspaper subcorpora describing refugees and asylum seekers. Such studies illustrate the usefulness of collocation comparisons between corpora and highlight the need for tailored statistical approaches.

In this paper, we will discuss a range of applications of a co-occurrence comparison method that we proposed to assess collocational behaviour across corpora (Wiegand et al., 2016). The aim of the present paper is to show how variations on the method can be applied to different linguistic questions that can be answered through the comparison of collocation. In various case studies we consider the influence of factors such as narration style, editorial influence and the date of publication; in each case we discuss how these factors can be explicitly considered or removed by using different variations of the method. Through the focus on such variations we illustrate the relevance of the technique to a wide range of research questions. While the paper focuses on the applicability and implications of the comparison of co-occurrences, we will provide pointers to the relevant underlying statistical theory, which is covered in detail in Wiegand et al. (2016). These methods borrow heavily from the discipline of meta-analysis as used in medical research (see e.g., Cooper, Hedges & Valentine, 2009). However, the application and interpretation of the methods are different when considered in the context of corpus linguistics. We will highlight these differences in the paper.

Method and corpora

Our CorporaCoCo method is a systematic statistical method to analyse the differences in the collocational behaviour of words across corpora (Wiegand et al., 2016). In this paper, we briefly explain the overall approach and highlight crucial statistical concepts, but the main focus is on the application and interpretation of the method and its variations. The CorporaCoCo method directly compares co-

occurrence counts for a set of node terms with all other words in the corpora. Fisher's exact test is used with a multiple test correction. The visualisations show an effect size with an associated confidence interval providing a clear overview of the results. Both the method and visualisation tools are made available in the CorporaCoCo R package (Hennessey et al., 2017), which we will also briefly introduce. Our approach to collocation comparison is applicable to any register and can be used for the conceptualisation of discourse more widely. For illustration, the paper uses case studies comparing novels by Charles Dickens to a set of other 19th century novels.

A simple comparison of collocation behaviour between corpora

We compare all 15 Dickens novels to a set of other 19th century novels. We are not interested in any corpus-internal variation but effectively 'average' the contents of each corpus. The question we are asking in our case study is: "what are the differences in the lexical patterns of body part nouns in Dickens's novels compared to the set of other 19th century novels?" This question is based on the importance of body part language in Dickens's fiction as described by Mahlberg (2013). In particular, we are looking for differences between the corpora and we are specifically not considering variation within each of the two corpora. Statisticians refer to this type of analysis as a fixed-effects model. A fixed-effects model is usually applied when an effect is expected to be uniform across the data. Here, we use the CorporaCoCo method specifically to take advantage of the property of the fixed-effects model to average variation within the corpora.

A comparison of collocations between corpora that considers the dispersion of the effect

If we want to acknowledge that there may be a difference in co-occurrence between the subcorpora (i.e. corpus-internal variation), a random-effects model can be used. The use of random-effects models has been put forward by some corpus linguists (e.g. Gries, 2015). Our use of random-effects models represents the lack of knowledge about the difference in co-occurrence between subcorpora as some kind of random effect. Crucially, the corpus-internal variation is no longer ignored or averaged out, and one of the outputs of the method is a measure of the dispersion of the effect across the subcorpora.

The level at which we consider partitioning into subcorpora, and so for which we can identify dispersion, is important and depends on the research question. In our case study we partition Dickens's novels into five subcorpora based on the periodicals in which they were originally serialised to examine editorial impact on the dispersion of co-occurrence patterns.

Conclusion

This paper extends our argumentation for comparing collocation across corpora, developed in Wiegand et al. (2016), to a variety of corpus linguistic applications using fixed- and random-effects models. Based on case studies comparing Dickens's novels against other 19th century novels we have shown that a simple fixed-effects

model, like the one distinguishing narration styles, can be very powerful. We also offer techniques applying the more complex random-effects models that allow us to consider dispersion of the co-occurrence effects, for example across different periodicals, whilst only making weak assumptions about the variation across subcorpora. While we use case studies to provide concrete examples of how our CorporaCoCo method can help to address specific research questions, the paper overall aims to make a convincing case for the need and the applicability of our proposed method.

References

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The Handbook of Research Synthesis and Meta-Analysis* (2nd ed.). New York: Russell Sage.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (pp. 1212–1248). Berlin: Walter de Gruyter.

Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: a corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press, 1996-2005. *Journal of English Linguistics, 36*(1), 5–38.

Gries, S. T. (2013). 50-something years of work on collocations: what is or should be next ... *International Journal of Corpus Linguistics*, *18*(1), 137–166.

Gries, S. T. (2015). The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora*, *10*(1), 95–125.

Hennessey, A., Wiegand, V., Mahlberg, M., Tench, C. R., & Lentin, J. (2017). CorporaCoCo: Corpora Co-Occurrence Comparison (Version 1.0-2). Retrieved from https://cran.r-project.org/package=CorporaCoCo

Mahlberg, M. (2013). *Corpus Stylistics and Dickens's Fiction*. London: Routledge.

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Sinclair, J. (Ed.). (1987). *Looking up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. London: Collins.

Wiegand, V., Hennessey, A., Tench, C. R., & Mahlberg, M. (2016). Comparing co-occurrences between corpora. Manuscript in preparation.