Corpus Linguistics, Language Data Science, and Computational Linguistics – building bridges or splitting apart?

Martin Schweinberger

Abstract

Computation has doubtlessly revolutionized our world and has become an integral part of all domains of everyday life – from research and finding information over ordering food and booking travels to dating and entertainment – and all this within merely one or two generations. The "digital" has impacted practices in corpus linguistics and has been offering novel pathways for research as we are now able to make use of vast amounts of electronically available data and complex modelling of language and linguistic behaviour (Anthony, 2020). The integration of computation and quantitative statistical means to understand linguistic processes has indeed allowed us to gain important insights into how language changes and cognitive processes underlying language use. It has further resulted in language data-based applications (search engines, machine translation, speech recognition, etc.) that would have been hard to imagine just two generations ago. When it comes to linguistic theory-building, especially the role of frequency – and thus usage-based approaches to understanding cognitive processes – have been rapidly gaining ground. This too would not have been possible without being able to scrutinize large amounts of natural language data.

Yet, this computational revolution has also sparked conflict and has been criticised – leading to arguments that corpus linguistics has become too methods- and data-focused while neglecting the more traditional focus on "linguistic description" (Larsson, Egbert & Biber, in press). This scepticism has raised awareness of potential problems and culminated in discussions about the appropriate use of digital methods – ultimately, the varying responses to linguistics becoming more data-sciency pose questions about the future direction of our field: should or will corpus linguistics in the end remain part of the humanities or could it shift towards engineering?

This talk intends to further this discussion and to address relevant questions relating to the relationship between corpus linguistics, language data science, and computational linguistics using selected case-studies and examples of emerging infrastructure projects. It argues for the need of training and infrastructures in computational skills and the integration of basic programming in linguistics programs, and it discusses how the computational revolution will continue to impact our field.

References

Anthony Laurence. (2020) Programming for Corpus Linguistics. In Magali Paquot and Stefan Thomas Gries (eds.), *A Practical Handbook of Corpus Linguistics*, 181-207. Springer.

Larsson, Tove, Egbert, Jesse, & Biber, Doug. (in press). On the status of statistical reporting versus linguistic description in corpus linguistics: A ten-year perspective. *Corpora* 17(1).