Pre-conference workshop:

Corpus pitfalls: dealing with messy data (and other traps for the unwary)

Convenors:

Mark Kaunisto (Tampere University),
Marco Schilk (Universität Hildesheim),
Jukka Tyrkkö (Linnæus University)

In a short but important paper published thirty years ago in ICAME Journal, Rissanen (1989) identified and discussed three potential problems that the use of corpora may present to unwitting scholars not necessarily closely familiar with the contents, structure and overall character of the corpora that they set out to examine. As solutions to the issues raised, Rissanen promoted the compilation of larger and more representative corpora, as well as getting acquainted with the data itself. Today we have access to corpora which are massive, compared to the ones available thirty years ago, which may widely improve representativeness. At the same time, paradoxically, these increases in corpus size make it considerably less likely that corpus users will have a clear understanding of the characteristics of many of the texts (or text types) included in the corpora. In other words, the sound advice of "knowing your data" becomes increasingly hard to follow when working with corpora that consist of billions of words from a huge number of different sources. It may therefore be argued that larger corpora have also brought about new kinds of problems (as noted, e.g., by Hiltunen, McVeigh & Säily 2017).

Finding unwanted items among search results is probably a very typical, almost everyday experience for many corpus linguists. Corpus compilers spend considerable effort regarding corpus annotation, markup, boilerplate removal, identification of duplicates or OCR errors. Similarly, scholars using corpora use increasingly refined methods when constructing elaborate query strings. Yet, achieving perfect precision and/or recall is still highly unlikely. We often need to exclude different kinds of search hits from further analyses, and the reasons for weeding out unwanted items can be varied. Occasionally the occurrence of false positives is mentioned in research articles, perhaps in a footnote, but it may also be the case that much of the clean-up of irrelevant items is done silently.

For example, finding a search term within a quotation in a corpus text might justifiably give rise to exclusion of a token from further analysis. In fact, quotations can constitute a significant part of many corpora even in terms of their word count, yet their role overall in corpora has received little attention (see e.g. Rissanen 1992; Kaunisto 2017). Another related issue concerns the inclusion of various levels of linguistic annotation in corpora, which are often accepted as given especially by less experienced corpus linguists, but which may at times be less than helpful (see, e.g., Sinclair 2004; Archer 2012). Furthermore, the dispersion of tokens across the corpora can be a significant factor when assessing search results (see e.g. Gries 2008).

There are undoubtedly many types of persistent problems and messiness in corpus data that seasoned scholars have encountered and know about, but which are seldom specifically addressed. Yet beginning corpus users might benefit from learning about what may be regarded as tacit knowledge in corpus linguistics, and even the more advanced scholar may encounter issues new to them that have been addressed earlier. This workshop intends to tap into this knowledge by inviting papers on the following topics:

- false positives found in corpora; how to find them or assess their frequency in a corpus?
- the significance of identifying different types of unwanted items; how to deal with them and what are the risks if they are not identified?
- problems associated with categories built into corpus design and various types of linguistic annotation in corpora; to what extent can these seemingly helpful features encourage uncritical thinking or guide corpus users research?

It deserves to be mentioned that problematic aspects may be detected in individual corpora, and observing such infelicities as well as dealing with them is without question necessary and useful as the aim of such observations is to advance corpus linguistic endeavours on the whole. However, instead of focusing on corpus-specific issues, this workshop welcomes papers that reflect on general issues or their own experiences of, and mistakes in, corpus compiling and corpus-based research. In the collegial spirit of ICAME, this workshop is not intended as a forum for highlighting mistakes or shortcomings in fellow scholars' work.

The estimated number of participants is 6-8, which means that the workshop would take roughly four to five hours. The conveners of the workshop are planning an edited volume based on the papers presented at the workshop.

Call for papers

Abstracts should be between 400 and 500 words in length (excluding references) and both full papers and work-in-progress reports are welcome. They should be sent via e-mail to Mark Kaunisto at mark.kaunisto@tuni.fi; the deadline for abstract submission is December 31, 2020. Notifications of acceptance will be sent out by mid-January 2021.

Appendix

Archer, Dawn. 2012. Corpus annotation: a welcome *addition* or an *interpretation too far*?, in *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources*, edited by Jukka Tyrkkö, Matti Kilpiö, Terttu Nevalainen & Matti Rissanen. Studies in Variation, Contacts and Change in English 10. Helsinki: eVarieng. Available online at http://www.helsinki.fi/varieng/series/volumes/10/ archer/

Gries, Stefan Th. 2008. "Dispersions and adjusted frequencies in corpora". *International Journal of Corpus Linguistics*, 13:4, 403–437.

Hiltunen, Turo, Joe McVeigh & Tanja Säily. 2017. "How to turn linguistic data into evidence?", in *Big and Rich Data in English Corpus Linguistics: Methods and Explorations*, edited by Turo Hiltunen, Joe McVeigh & Tanja Säily. Studies in Variation, Contacts and Change in English 19. Helsinki: eVarieng. Available online at http://www.helsinki.fi/varieng/series/volumes/19/ introduction.html.

Kaunisto, Mark. 2017. "Multilingualism and quotations from a corpus-linguistic perspective: a case study of Samuel Taylor Coleridge's *Biographia Literaria*", in *Challenging the Myth of Monolingual Corpora*, edited by Arja Nurmi, Tanja Rütten, and Päivi Pahta, 220-238. Leiden: Brill.

Rissanen, Matti. 1989. "Three problems connected with the use of diachronic corpora". *ICAME Journal* 13: 16-19.

Rissanen, Matti. 1992. "The diachronic corpus as a window to the history of English", in *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, edited by Jan Svartvik, 185-205. Berlin and New York: Mouton de Gruyter.

Sinclair, John. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.