Corpus Tools Today and Tomorrow: Incorporating User-Defined Annotations

Nicholas Smith,[1] Sebastian Hoffmann[1]
and Paul Rayson[1]

Abstract

Today's corpus tools offer the user a wide range of features which greatly facilitate the linguistic analysis of large amounts of authentic language data (e.g. frequency distributions, collocations, keywords, etc.). However, these tools typically fail to address one fundamental need of the linguist, viz. to add interpretive information to a query result by coding individual concordance lines for structural, functional and discoursal features that are deemed relevant to a fuller understanding of the phenomenon under investigation. For example, a user who has retrieved a list of BE + past participle constructions may wish to indicate which instances are – or are not – true passives; which have an agent phrase, or a subject representing "given" information, in the co-text; and which are used as part of an apology or indirectness strategy. Apart from marking such categories on a print-out of the results, one of the standard solutions for this task is to export the concordance to a general-purpose database or spreadsheet program, which permits multiple levels of user-specified annotation and offers advanced filtering and arithmetic functions that can help uncover patterns of behaviour (see e.g. Kirk 1994). However, a major drawback of this type of approach is that the link to the original source text is severed and that the advanced functions of the corpus tool – collocations, keywords, n-grams, etc. – cannot in turn be applied to the manually post-processed set of results.

Our paper has two aims: first, we will take stock of the currently available tools and strategies for manual analysis of a corpus query result, outlining both the advantages and drawbacks of the various options. More importantly, however, we wish to draw up a set of desiderata for the incorporation of flexible encoding features into future corpus tools which will then more adequately meet the needs of researchers in their analysis of linguistic data.

References

Kirk, J. 1994. "Corpus–Concordance–Database–VARBRUL." *Literary and Linguistic Computing.* 9(4): 259–66.

[1] Lancaster University
e-mail: paul@comp.lancs.ac.uk